How to assess progress in competency-based postgraduate medical training
For reasons of consistency within this thesis, some terms have been standardized throughout the text. As a consequence the text may differ in this respect from the articles that have been published.

The studies presented in this thesis have been performed at the Department of Primary and Community Care and the Scientific Institute for Quality of Healthcare (IQ healthcare). These are part of the Radboud Institute Health Science (RIHS), one of the approved research institutes of the Radboud University Nijmegen Medical Centre.

Financial support by the Scientific Institute for Quality of Healthcare (IQ healthcare) for the publication of this thesis is gratefully acknowledged.

ISBN: 9789462792753

Nijmegen, 2014
How to assess progress in competency-based postgraduate medical training

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. Th.L.M. Engelen,
volgens besluit van het college van decanen
in het openbaar te verdedigen op

maandag 27 oktober 2014 om 14.30 uur precies

door Frederik Tromp

geboren te Haarlem
op 27 mei 1957
Promotor: Prof. dr. M.J.F.J. Vernooij-Dassen

Copromotoren: Dr. A.W.M. Kramer  
              Dr. B.J.A.M. Bottema

Manuscriptcommissie: Prof. dr. A.M. van Dulmen  
                       Prof. dr. Th.J. ten Cate (UMC Utrecht)  
                       Dr. C.R.M.G. Fluit
## Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1</td>
<td>Introduction</td>
<td>7</td>
</tr>
<tr>
<td>Chapter 2</td>
<td>Behavioral elements of professionalism: Assessment of a fundamental concept in medical care</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Published in Medical Teacher 2010; 32: e161-e169</td>
<td></td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Assessment of CanMEDS roles in postgraduate training: the validation of the Compass</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>Published in Patient Education &amp; Counseling 2012; 89: 199-204</td>
<td></td>
</tr>
<tr>
<td>Chapter 4</td>
<td>How do trainers form an overall judgment of trainees performance and progress</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Submitted</td>
<td></td>
</tr>
<tr>
<td>Chapter 5</td>
<td>Development of a multi-method selection procedure for postgraduate training based on the CanMEDS in the context of GP training</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>Submitted</td>
<td></td>
</tr>
<tr>
<td>Chapter 6</td>
<td>Summary and general discussion</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>Samenvatting</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>Dankwoord</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td>Curriculum Vitae</td>
<td>104</td>
</tr>
</tbody>
</table>
Chapter 1

General introduction
Until recently, postgraduate medical education was mainly characterized as learning by doing under supervision. The characteristics of the learning experience included participation, unstandardized workplace experience, supervision and expert judgment. Because of societal demands for increased transparency, competency-based medical education (CBE) seeks to make the outcome of training more explicit by describing what competencies have to be mastered at the end of training. Competencies have been described as the "habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in daily practice for the benefit of the individual and the community being served". This definition emphasizes the operational nature of competencies. Competencies do not represent static knowledge and the memorization of facts but rather the behaviours and nature of actions that are carried out while utilizing one's knowledge. Competency refers to both what one knows as well as how one acts when applying that knowledge. This definition also underlines the importance of learning in practice (workplace-based learning). Moreover, CBE aims to widen the scope of training by including a broader spectrum of learning domains in order to prepare future doctors for the challenges they will meet due to societal developments as aging, a shift to primary care, rapid knowledge development and financial constraints. Finally, CBE underlines the importance of learner-centred, personalized training.

Traditional assessment methods have generally emphasized knowledge and skills acquisition, with minimal formal assessment of performance. With the introduction of competency-based education, it becomes important to devise parallel strategies to assess these competencies. These strategies include both the content and the process of assessment. For the content the broader scope of CBE means that not only clinical competencies (medical expertise, patient communication, professional behaviour) are assessed but also the more generic competencies as management, health advocacy, organization and collaboration. For the process the emphasis on performance in practice means that assessment procedures should focus on the workplace. Moreover, in order to guarantee learner-centred, personalized training, the attention should be more on guiding and monitoring the learner's professional development. The actors in the complex situational setting of workplace-based learning include the trainee, who must develop clinical and generic competencies, and the trainer, who guides and assesses this development. Instruction occurs in an actual patient care setting that cannot be fully structured or standardized for assessment purposes. For this reason, assessment programmes are recommended that are composed of different assessment methods. In such a programme, progress is determined after the careful aggregation and consideration of information that is obtained from a variety of sources and instruments and collected over a prolonged period of time. To stimulate learning behaviour of trainees, these assessment instruments should be capable of generating well-defined feedback.
An important step is the development and validation of assessment instruments that correspond with the principles of competency-based medical education. Therefore, we focus in this study on the development and validation of assessment procedures that can evaluate performance at the workplace, covering a broad scope of competencies and giving insight in learner’s development. We perform our study in the context of the Dutch postgraduate training in general practice that has embraced the principles of competency-based education. This introductory chapter sketches the context of Dutch postgraduate training in general practice, the theoretical background, the central aim of this thesis and the specific aims resulting from the historical educational developments in Dutch GP-training and, finally, the outline of the next chapters.

The context of Dutch postgraduate training in general practice
Dutch general practice trainees work independently in a general practice under supervision of a general practitioner (GP) trainer. Postgraduate training for general practice is offered by the Departments of Family Medicine of the eight university medical centres in the form of a dual-track programme that lasts for approximately three years, depending on the trainee’s prior experience and personal circumstances. Because all of the programmes must comply with the national curriculum framework, they are highly comparable; each programme is composed of two years of training in general practice interspersed with one year in hospital and community services, during which the trainees participate in health care and learn in interaction with their work environment. In addition to the workplace-based programme, each department offers a more academically oriented ‘day release programme’ in which the trainees weekly reflect on their learning in practice, learn about selected topics, and evaluate and direct their own learning in one-to-one progress interviews with their teacher.

Since 2006 the training has adopted a competency-based approach. The curriculum is now based on the CanMEDS competencies, which are adapted to the specific needs of the specialty of general practice. During the years elements of theoretical models and frameworks have been used to develop the assessment. To assess the seven CanMEDS-competencies on an aggregated level the Competency Assessment List (Compass) is used.

Theoretical background
General view
Viewpoints are evolving from the assessment of learning to the assessment for learning. It has been argued that learning is the key purpose of assessment: assessment drives learning. Feedback resulting from assessment is a core component of this so-called formative assessment, which is central to learning and is at ‘the heart of medical education’. Formative assessment is not merely intended to assign grades to trainee performance at designated points in the curriculum, it is part of the instructional process to support and enhance learning. Competency-based training is self-directed, individualized to the learning needs of the trainee and aimed to
stimulate life-long learning\textsuperscript{11}. Also, teaching is considered a collaborative process in which responsibility is shared between trainer and trainee. This collaboration requires the trainee to be an active participant in determining a learning plan. Critical skills required to the trainee are reflection and self-assessment\textsuperscript{12}. These viewpoints shape the assessment tools and procedures in competency-based medical education. They are based on various models and frameworks described below.

1. Theoretical models and frameworks guiding the assessment of competency-based medical education

As a first step in the development and validation of competency-based assessment procedures we have studied the literature on assessment and competency-based education. In this paragraph we will discuss the assessment literature based on the theoretical model of Miller's classification of clinical performance, the frameworks of workplace-based assessment as proposed by Pangaro and Ten Cate and the conceptual model of programmatic approach of van der Vleuten. Moreover, we will describe the CanMEDS-framework that provides a guide for the content of CBE and is used in many countries today, including the Netherlands. All these models and frameworks describe the process of assessment, except for the CanMEDS, which describes the content of what should be assessed.

Classification of clinical performance

In the traditional medical education and assessment, knowledge tests played the most important role. In 1990, George Miller presented a classification of clinical performance and its assessment, using a pyramid to illustrate different levels and how best to assess them (figure 1)\textsuperscript{13}. The base of the pyramid represents knowledge and the extent to which the learner understands what is required for competent professional performance. Knowledge can be assessed reliably and validly using open and multiple choice questions in addition to oral assessments\textsuperscript{13,14}. However, a learner must also know how to apply knowledge. Assessment at this level is similar to knowledge assessment but is more closely connected to a professional context, for example, by use of patient scenarios and vignettes\textsuperscript{14,15}. Next, the 'shows how' level moves beyond the merely cognitive level to actual behaviour. What are learners actually doing when they apply their knowledge? This can be assessed in simulated and standardized situations, such as Objective Structured Clinical Examinations (OSCE). Finally, the top of the pyramid focuses on what a professional actually is doing when functioning independently in the clinical workplace\textsuperscript{13,14}. As learning in practice is the heart of competency-based education, the focus of assessment in CBE should be situated at the very top of the pyramid (does-level).
Frameworks of workplace based assessment

In contrast with the traditional formal assessments (knows-shows how), competency-based assessment focuses on assessment at the workplace. Although the observation and assessment of learner's performance in real-life professional settings has always been a cornerstone of (post) graduate medical education, it was until recently not formalized and examined. Frameworks for workplace-based assessment are needed to give clarity to trainers and learners about what is to be evaluated in practice. What kind of observations or assessments are useful, and how are these analysed and compared to a standard of what is to be achieved by the learner. Frameworks contain a group of ideas or categories to reflect the educational goals against which a learner's level of competence or progress is measured. They enable trainers to determine to what extent learners are ready for advancement, meaning that the desired competence has been attained. Therefore, frameworks should provide both a valid mental model of competence and also terms to describe successful performance, either at the end of the training or as milestones during the curriculum. As a consequence, frameworks drive learning by providing learners a guide for what they are expected to learn.

In assessment of postgraduate medical trainees on the upper Miller level, trainers may use their own experiences as a benchmark. To clarify the standard of what is to be achieved by the trainee, several frameworks of workplace-based assessment can be described. Pangaro and Ten Cate make a distinction between analytic, synthetic and developmental frameworks for assessment.

Analytic frameworks deconstruct competence into individual parts to evaluate each part separately, for example knowledge, skills and attitudes. This should facilitate measurement.
Chapter 1

Synthetic frameworks view competence holistically, focusing on evaluation of the performance in real-world activities. This approach is essentially integrative. The grounding question is: what activity can be entrusted to a trainee, once sufficient competence has been reached? Such tasks are named "entrustable professional activities" (EPA)\(^{18}\). In EPAs multiple competencies are required and must be brought together (synthesized).

Developmental frameworks focus on stages or milestones in the progression toward competence. The assumption in the developmental model is that there are stages of progression in a logical order\(^{19}\). Carraccio and colleagues provided terms for what progress in medical expertise looks like, for example from 'novice' to 'advanced beginner' to 'expert'.

**Programmatic approach of assessment**

Competency-based medical education relies on continuous, comprehensive and elaborate assessment and feedback systems\(^{20}\). To ensure that development and acquisition of necessary competencies is assessed in an appropriate and comprehensive manner, Van der Vleut\(\text{e}^\text{t}\) et al. advocated a programmatic approach\(^{5}\). In an assessment programme, decisions are not taken on the basis of a single assessment but are instead taken after the careful aggregation and consideration of information from a variety of sources that are collected over a prolonged period of time. At the end of a training period, all the information that has been collected by the trainer is assessed in an intermediate progress evaluation. Judgments are made by aggregating the available information across data points. The results should be discussed between trainer and learner, and the feedback should be incorporated in learning plans. This way assessment drives learning. An assessment programme provides an overview of what is to be measured, thus counteracting the pitfall of overemphasizing easy to measure elements, such as for instance factual knowledge\(^{21}\).

**The CanMEDS-framework**

Until now, we discussed frameworks that describe the process of assessment, how to assess, the next framework describes the content, what should be assessed. The framework developed by the Canadian Medical Education Directives for Specialists (CanMEDS) 2000\(^{22}\) guides the construction of curricula in many countries today. In the context of a rapidly changing health care environment, it was felt that the roles and abilities required of physicians needed to be further defined and explored. The Canadian College identified patient consumerism, government regulations, financial constraints, medical information on the Internet, litigation, technology and the explosion of medical knowledge as forces that change the nature of health care delivery. Within this context, the question arose: "How can we best prepare physicians to be effective in this environment and truly meet the needs of their patients?" This question led to a competency-based restructuring of the curricula for medical education broadening the scope of training by including a wider spectrum of learning domains. The CanMEDS-framework stresses the central
role of the medical expert, but it adds six other (non-medical) domains that are also considered essential: communicator, collaborator, manager, health advocate, scholar and professional (figure 2)\textsuperscript{22,23}.

\begin{figure}[h]
\centering
\includegraphics[width=0.6\textwidth]{canmeds_model.png}
\caption{The CanMEDS model}
\end{figure}

2. What is known in the literature about competency-based assessment procedures in medical education?

In addition to the theoretical frameworks we have searched the literature for empirical studies into competency-based assessment procedures. Since assessment in CBE predominantly takes place in the workplace, we have examined the literature on workplace-based assessment. Several assessment tools have been developed for use in the workplace and there is evidence of their reliability and validity\textsuperscript{6,24}. Tools that are described most frequently are the mini-Clinical Evaluation Exercise (mini-CEX) and the direct observation of procedural skills (DOPS). Research on assessment in medical education has strongly emphasized individual instruments that measure performance during a single patient encounter\textsuperscript{25}. Although single-point assessments are useful in generating insight in specific performance, feedback and support, they do not provide a sufficiently clear picture of the trainee's progress and overall performance\textsuperscript{6,7,25}.

To overcome the problem of single-point assessments Multisource Feedback (MSF) was developed. It is a method in which performance is assessed over a longer period of time: instead
of judging individual encounters, assessors rely on their exposure to the trainee's work for an extended period of time. MSF includes often a self-assessment and assessments from a range of others who are in a position to give relevant judgments of one or more competencies such as communication, collaboration or management. These may include peers, trainers, patients, other health-care workers, etc. With sufficient sampling across assessors these, often global, performance evaluations can give reliable judgments about the topics that are assessed.

Although single encounter assessments and MSF give information about performance, this information has to be integrated and aggregated with informal performance information about all competencies to come to an overall judgment of performance and progress. One of these aggregation methods that samples performance across a longer period of time is the portfolio. Portfolios can be classified in terms of the functions they have: monitoring and planning, coaching and reflection, and assessment. They can play a minor part in a larger assessment programme or they can be the main method to aggregate and evaluate all assessments at the 'does' level. Portfolios tend to work best if functions are combined and if they perform a central function in guiding learning, in coaching and in monitoring longitudinal competency development. In a portfolio a learner gathers the performance data. We did not find instruments that are helpful to the trainer to integrate and aggregate these longitudinally obtained data.

3. The role of the users in competency-based assessment

Our final step was to search the literature about the users in workplace-based assessment. Assessment at the does level (in the workplace) is less structured and organized and more subjective than in the other layers of Miller's pyramid. Therefore, the actors (learners and trainers) play an important role in if and how the assessment is performed. The literature shows that their role is even more vital than the instruments themselves. Trainers' expertise in using the instrument, their personal ideas about assessment, the extent to which they take the assessment and instruments seriously and the time they can spend on it, they all determine whether or not the assessment is well performed. Also, the learners play an important role. The extent to which they take the assessment seriously and are open for feedback, how active they are in asking for observation and feedback and how they use the feedback to improve performance, they all also determine the effectiveness of the assessment. Ensuring that the users have a proper understanding of their roles requires training and expertise development.
Aim of this thesis

Central aim
The theoretical models in the literature show that in competency-based assessment the focus should be on workplace-based assessment\textsuperscript{13}, on frameworks to give clarity to trainers and trainees about what is to be evaluated in practice\textsuperscript{17}, on programmatic assessment to guide and monitor professional development\textsuperscript{5} and on the content\textsuperscript{13}. Several validated instruments are available for single-encounter assessment. With the validated multisource feedback procedure trainees can be globally assessed by co-workers. A portfolio is helpful in gathering of and an easily access to information. However, there is no evidence on how to combine all this information into an aggregated assessment of overall professional performance and progress. Therefore the central aim of this study is to develop and validate an instrument that can support this process in order to get an appropriate picture of the professional growth on all competencies and to generate meaningful feedback to guide learning.

Our focus will be on the content of the instrument (does level, framework), on the reliability, on the feasibility and on the process (how do the trainers use the instrument). We conducted this study in the ’real world’ of GP-training and had to adapt our study to the ongoing educational developments that were implemented. The results from our first study were used in the next study and so on, giving our studies ecological validity. Ecological validity is the degree to which the behaviours observed and recorded in a study reflect the behaviours that actually occur in the real world\textsuperscript{31}. In the following sections we shall sketch the historical developments of competency-based assessment in GP-training in the Netherlands and present the resulting specific aims of each study.

Specific aims per study resulting from the historical educational developments
The theoretical models and frameworks discussed in the theoretical background are incorporated in the assessment programme of Dutch GP training. The assessment follows a programmatic approach. The focus is on the assessment of progress of performance in daily practice. Decisions are not taken on the basis of a single assessment but are instead taken after the careful consideration of information from a variety of sources that are collected over a prolonged period of time. The Competency Assessment List is used to assess the seven CanMEDS-competencies separately (analytic approach). The competencies themselves are not divided into separate parts, and are assessed in practice (synthetic approach). The reference point of assessment is the standard of performance required upon completion of the training.

Study 1
Although CBE was nationally introduced in 2006, Dutch GP-training acknowledged already earlier the importance of a broader scope of training content. The Dutch Basic Job Description guided
the training programme. It includes four domains: 1) providing care, problem solving and care provision adapted to the situation in general practice, 2) categories of patients, complaints and diseases, 3) ability to organize and manage a general practice, and 4) personal functioning as a GP (corresponding with professionalism). However, tools to assess these domains in practice were lacking. As professionalism is considered to be fundamental in medical care, the development of a workplace-based assessment instrument was initiated in 2002 (the Nijmegen Professionalism Scale)\textsuperscript{32}. A first step was to describe and validate the content. After reviewing the literature professionalism was conceptualized as a broad competency in which four themes could be distinguished: professionalism towards the patient, professionalism towards other professionals, professionalism towards society and professionalism towards oneself\textsuperscript{33,34}. Items of the Nijmegen Professionalism Scale were defined as examples of professional behaviour as they occur during everyday practice. Item development arose from an extensive review of the literature. The developers looked for specific, prescriptive examples of professional behaviour. A Delphi study generated additional items. In two rounds experts in the field of Professionalism were invited to formulate behavioural examples.

Our first study started with a further validation of the Nijmegen Professionalism Scale (NPS) by examining its construct validity and reliability. The primary aim of the NPS, is formative assessment. It allows GP trainers to systematically provide feedback about the professional behaviour of their GP trainees. The GP trainers have to complete the instrument every three months to evaluate their GP trainees. The results are used in a progress meeting. These evaluations are thus not single-moment assessments, but cover a three-month period of multiple observations. The goal of the progress meeting is to formulate "professional behaviour learning points" for the GP trainee. These learning points are issues within professional behaviour selected to be improved systematically by setting goals for the future. In generating learning points, the GP trainees are encouraged to reflect on their strong and weak points in professional behaviour.

Study 2 and 3
After CBE was officially introduced in Dutch specialty training, including GP-training, the CanMEDS framework was adopted and adapted to the specific requirements of the Dutch situation. The seven competency fields the Central College of Medical Specialties defined were: medical expertise, communication, collaboration, science and education, social accountability, management and professionalism\textsuperscript{35}. It was up to each specialty to describe the specific content of the competencies. It was also decided that with this transformation new strategies and instruments had to be developed to assess these competencies. For the GP-specialty it was decided to develop, as a first step in this transformation, a comprehensive, longitudinal, workplace-based assessment procedure. This procedure includes an instrument that assesses all seven competencies in GP-training: the Competency Assessment List (Compass). The trainees,
the trainers and the teachers of the day release course have to complete the Compass every three months during all three years of training. The aim is to assess progress of performance of all competencies, to generate feedback and to set new learning goals. The completed Compass is input for the three-monthly progress meetings. The development of the Compass was accomplished by consensus among experts in general practice and GP-training of all eight training institutes. Elements of the Nijmegen Professionalism Scale are used for the Compass. Due to its broad conceptualization the competency Professionalism showed much overlap with the other six competencies. For instance, many elements of Professionalism towards other professionals can be found in the competency Collaboration, and elements of Professionalism towards society can be found in the competency Social Accountability. Not only behavioural elements of the Nijmegen Professionalism Scale were included in the Compass, also the format of longitudinally assessment was applied. Both instruments follow a programmatic approach and guide the trainee towards the standard expected upon completion of the training. Decisions are not taken on the basis of a single assessment but are instead taken after the careful consideration of information from a variety of sources that are collected over a prolonged period of time.

Because of this development, the studies 2 and 3 focused on the Compass. We did a validation study into the content, reliability, feasibility and responsiveness of the instrument (study 2). Moreover, we were interested in how the trainers use the instrument. How do they form their overall judgments of trainee’s performance and progress on all seven competencies using the Compass.

STUDY 4
The last development we discuss in this thesis is the construction of a new competency-based selection procedure for admission to GP-training that we also consider as a baseline assessment to guide personalized learning from the very beginning of the training. During our study, research showed that the usual selection procedure had several flaws. This procedure was endorsed nationally but conducted locally. It was found that despite the legislation, different standards were used in different institutes. As a consequence, the department itself was the main predictor of being admitted and not the characteristics of candidates. Viewing the results of their study, the authors expressed their doubts about the fairness of the selection procedure and suggested that the current method should be reconsidered. A new procedure should be able to identify competent candidates and in addition, should provide information on individual competencies at the beginning of the training. The current literature on selection advocates to construct a selection procedure based on a job analysis. In our view the description of the seven competencies for the GP-training are comparable to a job analysis. The seven competencies listed in the Compass were the starting point of our job analysis. These competencies provide insight in what is needed to be allowed for independent practice as a GP. Moreover, starting trainees form a heterogeneous group, all individuals having different levels of competence due to
differing earlier working experiences. By assessing the same competencies as in training, trainers and trainees are able to use the results of the selection procedure to identify candidates' future development needs. With the aid of this feedback, future trainees may be able to formulate learning goals to remedy any potential shortcomings during the earliest stages of the training. Consequently, a selection procedure based on competencies as listed in the Compass can be seen as a baseline assessment, identifying shortcomings in one or more competencies.

Therefore, the fourth study of this thesis describes the development of a selection procedure that we constructed based on the CanMEDS competencies as listed in the Compass and that can serve as a baseline assessment from the very beginning of the training. Although the content of the procedure is guided by the Dutch GP-training context, the process of the development of the procedure is general and might be applicable for other competency-based specialties.

**Thesis outline**

Chapter 2 focuses on the competency Professionalism. The instrument discussed in this chapter is the Nijmegen Professionalism Scale. The construct validity and reliability of the Nijmegen Professionalism Scale is examined.

In chapter 3, we describe a study into the comprehensive progress assessment of the seven CanMEDS competencies as listed in the Compass. We did a validation study into the content, reliability, feasibility and responsiveness of the instrument.

In chapter 4, we present a qualitative study into how trainers form their overall judgments of trainees' performance and progress on all seven competencies using the Compass by examining the written considerations trainers have given for their judgments.

We describe in chapter 5 the development of a selection procedure based on the seven competencies as listed in the Compass. The procedure should be suitable for a baseline assessment at the very beginning of training.

Finally, in chapter 6, we present a summary of the main findings and the general discussion, in which we synthesize the findings of the studies described in this thesis. We also critically reflect on how to interpret these findings and what they add to the existing literature. We also consider the strengths, weaknesses and implications for future research and practice.

This thesis concludes with a summary in Dutch.
References


Chapter 2

Behavioural elements of professionalism: Assessment of a fundamental concept in medical care

F. Tromp
M.J.F.J. Vernooij-Dassen
A.W.M. Kramer
R.P.T.M. Grol
B.J.A.M. Bottema

Medical Teacher 2010;32:e161-e169.
Abstract

Background: The Nijmegen Professionalism Scale, an instrument for assessing professional behaviour of general practitioner (GP) trainees, consists of four domains: professional behaviour towards patients, other professionals, society and oneself. The purpose of the instrument is to provide formative feedback.

Aim: The aim of this study was to examine the psychometric properties of the Nijmegen Professionalism Scale.

Methods: Both GP trainers and their GP trainees participated. Factor analysis was conducted for each domain. Factor structures of trainee and trainer groups were compared. Measure of congruence used was Tucker's phi. Cronbach's α was used to establish reliability.

Results: Factor structures of the instrument used by GP trainers and trainees were similar. Two factors for each domain were found: domain 1, Respecting patient's interests and Professional distance; domain 2, Collaboration skills and Management skills; domain 3, Responsibility and Quality management; and domain 4, Reflection and learning and Dealing with emotions. Congruence measures were substantial (40.90). Reliability ranged from 0.78 to 0.95.

Conclusion: This study to validate the instrument represents one further step. To construct a sound validity argument, a much broader range of evidence is required. Nevertheless, this study shows that the Nijmegen Professionalism Scale is a reliable tool for assessing professional behaviour.
Background
Over the last decade, medical education has changed extensively. The focus has shifted from the acquisition of knowledge to the achievement of competence\(^1\). With the transformation into a competency-based programme, it is equally important that parallel strategies are chosen to assess these competencies. Competency-based teaching has been driven by competency frameworks such as the Accreditation Council for Graduate Medical Education and the American Board of Medical Specialties (ACGME/ABMS) competencies\(^2\), and the Canadian Medical Education Directives for Specialists (CanMEDS) 2000, issued by Canada Royal College of Physicians and Surgeons\(^3\). Competency-based education strives to assess the performance of residents. The basic essential elements consist of functional analysis of the occupational competencies, translation of these competencies into outcomes and assessment of the trainees' progress in these outcomes on the basis of their performance\(^4\). Assessments should be based on a set of clearly defined outcomes so that all parties concerned, including assessors and trainees, can make reasonably objective judgments about whether or not each trainee has achieved them.

One of the seven competencies that have been defined for the Dutch postgraduate training for family practice is professionalism. Professionalism is often cited as an essential part of medical performance and thus of medical training\(^5\)–\(^9\). However, professionalism has proved difficult to define\(^6\),\(^10\). Van de Camp et al.\(^11\) conducted a study to conceptualize professionalism, in which they reviewed the literature and proposed a multidimensional construct. The four domains they found within professionalism were: professional behaviour towards the patient, towards other professionals, towards society and towards oneself\(^11\),\(^12\). These domains provided the framework for the development of an instrument to evaluate the professional behaviour of general practitioner (GP) trainees. In the development of this instrument, it was decided to assess professionalism by focusing on behaviour rather than on traits\(^12\). Research showed that the key to valid assessment of professionalism lies in focusing in behaviour. Students do not identify themselves with abstract elements of professionalism, but they define professionalism in practical terms\(^13\),\(^14\). By framing professionalism in terms of behaviours rather than abstractions, we come much closer to a context bound, realistic framework for understanding professional behaviour\(^12\). Furthermore, observable behaviour is the appropriate basis of providing feedback\(^15\),\(^16\).

The primary aim of the Nijmegen Professionalism Scale is to generate formative assessment. It allows GP trainers to systematically provide feedback about the professional behaviour of their GP trainees. The GP trainers complete the instrument to evaluate their GP trainees every 3 months, and the GP trainees evaluate themselves with the Nijmegen Professionalism Scale. To become a GP in The Netherlands, 3 years of training are required after graduation as a medical doctor. During these 3 years, the GP trainee spends the equivalent of 4 days a week with a GP during the 1st and 3rd years. These periods are usually completed in different general practices,
where the GP then acts as a coach and a teacher. As the 1st and 3rd year are for practical training in a general practice, the 2nd year is dedicated to rotations through hospitals, clinics for chronically ill patients and psychiatric outpatient clinics. The assessments of the GP trainers and the self-assessments of the GP trainees are made independently. The results are discussed every 3 months in an interview of progress review. These evaluations are thus not one-time assessments, but cover a 3-month period of multiple observations. Both evaluations are compared in order to formulate 'professional behaviour learning points' for the GP trainee. These learning points are issues within professional behaviour selected to be improved systematically by setting goals for the future. In generating learning points, the GP trainees are encouraged to reflect on their strong and weak points in professional behaviour.

Instruments with good psychometric properties are needed for the evaluation of professional behaviour. Van de Camp et al. tested the content validity of the instrument in a qualitative study using the nominal group technique \(^{12}\), which consists of a very structured procedure to gather information from relevant experts \(^{17}\).

The goal of this study was to attain the best possible quality for the Nijmegen Professionalism Scale, which required the examination of its psychometric properties and further validation of the instrument by the assessment of its construct validity and reliability. Factor analyses were applied to determine its internal structure. The instrument is used by both GP trainees and trainers to evaluate professional behaviour. In this study, we compared the factor structure of these two groups. If the factor structure is the same for trainers and trainees using the instrument, then both GP trainers and trainees attach the same meaning to the construct of professionalism, and this will contribute to the validity. Cronbach’s \(\alpha\) was used to establish reliability.

**Method**

**Participants**

As a part of the curriculum, GP trainers and their GP trainees associated with the Department of Postgraduate Training for General Practice of the Radboud University Nijmegen Medical Centre complete the Nijmegen Professionalism Scale every 3 months. Permission was asked to analyze the data for this study. All 119 trainers and 119 trainees consented. Due to practical reasons, we only made use of the data of one single 3-month period, dated from September until November 2005.

**Measures**

The Nijmegen Professionalism Scale is an instrument with 106 items, each representing an element of professional behaviour. The instrument consists of four parts, each of which addresses a different domain within professionalism: professionalism towards the patient,
Professionalism towards other professionals, professionalism towards society and professionalism towards oneself. Each domain consists of separate scales (varying in number from four to nine) that measure different elements of professional behaviour. Following each item is a four-point Likert scale on which the participants can indicate how often a GP trainee exhibited the specified behaviour, ranging from 'seldom or never' (1) to 'always' (4).

As already mentioned, the GP trainers complete the instrument to evaluate their GP trainees every 3 months, and the GP trainees assess themselves with the Nijmegen Professionalism Scale. The trainers were instructed to discuss both evaluations in a one-on-one tutorial within 3 weeks of completing the instrument. Before using the instrument, trainers and trainees were informed in a brief training session about the primarily formative purpose of the instrument. Furthermore, a written manual was provided.

**Construct validity**

To examine the construct validity of the instrument, the four domains were analyzed separately, since each domain concerns a separate construct within professionalism. As a first step, a confirmatory factor analysis for each domain was performed to reproduce the original element structure. Whenever this analysis failed to replicate the original structure, an exploratory principal component analysis with varimax rotation was completed. Two criteria were used to determine the optimal number of factors to extract: the scree plot and interpretability of the factor loadings. Items were retained if they had a loading greater or equal than 0.40. Items with a factor loading less than 0.40 were discussed individually by researchers (FT, AK and RG) and GP trainers (MV and BB). If consensus was reached, items were retained or rejected. One of the criteria to reach consensus was face validity of the items. We felt that face validity was important because items with great face validity make the instrument conceptually clear and more straightforward for GP trainers and GP trainees to use.

**Construct equivalence**

In many cases, self-assessment and external assessment show only moderate agreement. Self-assessment, however, can be accurate under certain conditions, namely, when learners are expected to gather and interpret data about their performance and, at the same time, when they are required to reconcile their self-assessments with credible external evaluations. These conditions appear to be met with the Nijmegen Professionalism Scale. In testing the feasibility of the instrument, van de Camp et al. found very good agreement in the ratings of professional behaviour as observed by the GP trainer and the GP trainee. Consequently, both the data sets from the GP trainers and from the GP trainees were used and compared for this analysis.
Tucker's phi coefficients were computed for each factor. Phi values of 0.90 or more provide evidence of construct equivalence of both groups\textsuperscript{22}, which shows that both GP trainers and trainees attach the same meaning to the construct of professionalism.

**Internal consistency**
Cronbach's $\alpha$ was used to determine the internal consistency of items within each factor. We calculated Cronbach's $\alpha$ to provide additional evidence that the items within a factor were measuring the same underlying construct.

**Results**

**Sample**
The GP trainers and the GP trainees provided 116 lists that were eligible for inclusion. Three GP trainers and three GP trainees returned incomplete lists. The sample consisted of 60 1st-year and 56 3rd-year GP trainers and their GP trainees.

No indication of leniency, halo or ceiling effects were found as the scores ranged from 1 to 4 and showed sufficiently variance.

**Construct validity and equivalence**
Confirmatory factor analysis failed to replicate the original structure in all four domains. We, therefore, conducted an exploratory principal component analysis with varimax rotation. Examination of the scree plots after the principal component analysis indicated that there were two factors in each domain that best described the data.

**Domain 1: Professional behaviour towards the patient**
A two-factor solution was derived for both GP trainers and GP trainees. The results are shown in table 1.

<table>
<thead>
<tr>
<th>Items</th>
<th>Factors and item factor loadings</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GP trainer Factor 1*</td>
<td>GP trainer Factor 2*</td>
</tr>
<tr>
<td></td>
<td>Eigenvalue</td>
<td>Eigenvalue</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The GP trainee:</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Deals carefully with professional secrecy when talking to colleagues or acquaintances</td>
<td>0.44</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Deals correctly with legislative rules regarding informed consent</td>
<td>0.59</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Is able to bring up difficult subjects</td>
<td>0.51</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Respects the right of patients to inspect their medical records</td>
<td>0.41</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Does not give patients false hope</td>
<td>0.52</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Is able to show sympathy</td>
<td>0.49</td>
<td>0.48</td>
<td></td>
</tr>
</tbody>
</table>
Factors and item factor loadings

<table>
<thead>
<tr>
<th>Items</th>
<th>GP trainer Factor 1* Eigenvalue 5.42</th>
<th>GP trainer Factor 2* Eigenvalue 2.61</th>
<th>GP trainee Factor 1* Eigenvalue 5.91</th>
<th>GP trainee Factor 2* Eigenvalue 2.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 Takes care not to become part of the patient’s system</td>
<td></td>
<td>0.72</td>
<td></td>
<td>0.44</td>
</tr>
<tr>
<td>8 Takes care not to become too involved in the patient’s emotions</td>
<td></td>
<td>0.86</td>
<td></td>
<td>0.59</td>
</tr>
<tr>
<td>9 Takes care not to become too intimate</td>
<td></td>
<td>0.72</td>
<td></td>
<td>0.46</td>
</tr>
<tr>
<td>10 Takes patient’s opinions seriously</td>
<td>0.51</td>
<td></td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>11 Takes patients’ embarrassment, shyness, and reluctance into account</td>
<td>0.63</td>
<td></td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>12 During physical examinations, explains the aim of the procedures and what is expected of the patient</td>
<td>0.46</td>
<td></td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>13 Approaches patients with a different frame of reference (e.g., religion) openly</td>
<td>0.58</td>
<td></td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>14 Looks clean and tidy and dresses according to current norms</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>15 Adjusts language to communicate with patients with little education</td>
<td>0.47</td>
<td></td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>16 Takes sex-specific differences into account</td>
<td>0.43</td>
<td></td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>17 Is able to deal with cultural differences in the presentation and experience of complaints</td>
<td>0.53</td>
<td></td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>18 Takes care not to be influenced by patients of high social status</td>
<td></td>
<td>0.59</td>
<td></td>
<td>0.40</td>
</tr>
<tr>
<td>19 Is able to cope with the different expectations that patients have of their GP</td>
<td>0.40</td>
<td></td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>20 Involves the previous history of the patient in the provision of care</td>
<td>0.48</td>
<td></td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>21 Pays attention to the consequence of the treatment policy on the daily functioning of the patient</td>
<td>0.69</td>
<td></td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>22 Involves relevant aspects of the patient’s home and environment in the provision of care</td>
<td>0.55</td>
<td></td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>23 Retains insight into the medical history of patients in order to act proactively if necessary</td>
<td>0.62</td>
<td></td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>24 If necessary, takes action after life events</td>
<td>0.65</td>
<td></td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>25 Respects patients’ self-determination</td>
<td>0.51</td>
<td></td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>26 <em>Has difficulty taking decisions regarding diagnosis and treatment policy</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Total variance explained: 30.9% 30.8%

*Factors: 1 = respecting patient’s interests; 2 = professional distance. Factor loadings less than 0.40 not shown. Italicized items have been removed from the instrument.

The factor structure of the instrument used by the two groups is very similar. The only item that loads on a different factor in the two groups is ‘does not give patient false hope’.

The first factor was labeled respecting patient’s interests, since it comprised such behaviours as showing sympathy, adjusting language to communicate with patients with little education, taking gender-specific differences into account and dealing correctly with legislative rules. The second factor was labeled professional distance, since its items concerned such behaviours as ‘taking care not to become too involved in the emotions of the patient’ and ‘not becoming too intimate’.
We discussed the item 'does not give the patient false hope' and reached the consensus that it should be assigned to the first factor respecting patient's interests, following the structure found by the GP trainees. This decision was based on face validity.

Almost all items had a factor loading of at least 0.40, except the items: 'looks clean and tidy and dresses according to current norms' and 'has difficulty taking decisions regarding diagnosis and treatment policy'. We weighed the removal of these items against their educational significance. In our view, the educational significance of the item 'looks clean and tidy and dresses according to current norms' is considerable. Educators informed us that this item helped them raise an otherwise very difficult subject. The item was, therefore, retained despite its low factor loading and assigned according to its highest loading (0.30 on factor 1). The other item, 'has difficulty taking decisions regarding diagnosis and treatment policy', was removed from the list. In our judgment, this item did not fit in either of the two factors. The two factors yielded Tucker’s phi values of 0.95 and 0.94.

**Domain 2: Professional behaviour towards other professionals**

Here, also a two-factor solution was derived for both GP trainers and GP trainees. The factor loading matrix is shown in table 2.

| Table 2. Factor loading matrix in domain 2: professional behaviour towards other professionals (n=116) |
|---------------------------------------------------------------|---------------------------------------------------------------|---------------------------------------------------------------|---------------------------------------------------------------|---------------------------------------------------------------|
| Items                                                                 | The GP trainee:                                                                 | Factors and item factor loadings                                                                 |                                                                 |
| 1                                                                 | Is able to mediate with other care providers in the interests of the patient | GP trainer Factor 1* 0.48 | GP trainer Factor 2* 0.51 | GP trainee Factor 1* 0.48 | GP trainee Factor 2* 0.51 |
| 2                                                                 | Consults other care providers with targeted questions               | GP trainer Factor 1* 0.43 | GP trainer Factor 2* 0.53 | GP trainee Factor 1* 0.48 | GP trainee Factor 2* 0.48 |
| 3                                                                 | Discusses bottlenecks in cooperation with others directly           | GP trainer Factor 1* 0.58 | GP trainee Factor 1* 0.57 | GP trainee Factor 2* 0.50 |
| 4                                                                 | Complies with multidisciplinary working agreements                 | GP trainer Factor 1* 0.63 | GP trainee Factor 1* 0.57 | GP trainee Factor 2* 0.44 |
| 5                                                                 | Ensures structured information transfer with other care providers   | GP trainer Factor 1* 0.57 | GP trainee Factor 1* 0.65 | GP trainee Factor 2* 0.57 |
| 6                                                                 | Deals correctly with targeted questions from other care providers   | GP trainer Factor 1* 0.77 | GP trainee Factor 1* 0.65 | GP trainee Factor 2* 0.57 |
| 7                                                                 | Is able to write a good referral letter with direct questions       | GP trainer Factor 1* 0.47 | GP trainee Factor 1* 0.65 | GP trainee Factor 2* 0.57 |
| 8                                                                 | Is able to discuss a difference of opinion with a specialist directly | GP trainer Factor 1* 0.65 | GP trainee Factor 1* 0.65 | GP trainee Factor 2* 0.57 |
| 9                                                                 | Is able to manage the mutual demarcation of tasks between GP and specialists | GP trainer Factor 1* 0.49 | GP trainee Factor 1* 0.49 | GP trainee Factor 2* 0.57 |
| 10                                                                | Is able to influence specialist care (e.g., during consultation at hospital visits) | GP trainer Factor 1* 0.48 | GP trainee Factor 1* 0.48 | GP trainee Factor 2* 0.57 |
| 11                                                                | Ensures coherence in first and second line medical care             | GP trainer Factor 1* 0.40 | GP trainee Factor 1* 0.40 | GP trainee Factor 2* 0.57 |
| 12                                                                | Is able to motivate support personnel                              | GP trainer Factor 1* 0.58 | GP trainee Factor 1* 0.58 | GP trainee Factor 2* 0.61 |
| 13                                                                | *Chooses the correct time and place for comments about functioning* | GP trainer Factor 1* 0.58 | GP trainee Factor 1* 0.58 | GP trainee Factor 2* 0.61 |
Factors and item factor loadings

<table>
<thead>
<tr>
<th>Items</th>
<th>GP trainer Factor 1*</th>
<th>GP trainer Factor 2*</th>
<th>GP trainee Factor 1*</th>
<th>GP trainee Factor 2*</th>
</tr>
</thead>
<tbody>
<tr>
<td>14 Conducts structural consultations with support personnel</td>
<td>0.56</td>
<td>0.61</td>
<td>0.56</td>
<td>0.64</td>
</tr>
<tr>
<td>15 Makes clear agreements with support personnel</td>
<td>0.77</td>
<td>0.64</td>
<td>0.63</td>
<td>0.53</td>
</tr>
<tr>
<td>17 Is able to distinguish between professional and personal</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>interests in negotiations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 Is able to conduct job evaluations</td>
<td>0.63</td>
<td>0.60</td>
<td>0.63</td>
<td>0.56</td>
</tr>
<tr>
<td>21 Is able to help seek solutions if a colleague has too</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>high a workload</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22 Is able to provide emotional support for colleagues</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>23 Transfers services correctly</td>
<td>0.69</td>
<td>0.62</td>
<td>0.62</td>
<td>0.59</td>
</tr>
<tr>
<td>24 Shirks tasks</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Total variance explained</td>
<td>37.9%</td>
<td>40.4%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Factors: 1 = collaboration skills; 2 = management skills
Factor loadings less than 0.40 not shown. Italicized items have been removed from the instrument.

Again, the factor structure of the instrument used by both groups looks approximately the same. The item 'is able to manage the mutual demarcation of tasks between GP and specialists' loads on different factors. The item 'chooses the correct time and place for comments about functioning' has greater factor loadings in the trainee group, but it loads on both factors. The items 'conducts structural consultations with support personnel', 'is able to provide emotional support for colleagues' and 'shirks tasks' display higher factor loadings in the trainee group.

The items of the first factor included such behaviours as 'complying with multidisciplinary working agreements' and 'being able to motivate support personnel'. These behaviours were considered relevant to the relational part of collaboration with other healthcare workers; this factor was, therefore, interpreted as collaboration skills. The second factor included items related to management, such as 'being able to take policy decisions' and 'dealing constructively with conflicts'. This factor was labeled management skills. Five items (table 2) had a loading of less than 0.40 in the trainer group. These items proved to be of little educational significance according to the trainers, as these behaviours were seldom observed in practice. The items were, therefore, removed from the list. Tucker's phi values of the two factors were computed as 0.90 and 0.96.

**Domain 3: Professional behaviour towards society**

Table 3 shows the results. Two factors were determined. There are no considerable differences in the factor structure of the instrument used by the GP trainers and the GP trainees. Two items 'has perceptions about how form can be given to means of contact (telephone services, diabetes,
surgery hours, etc.)' and 'is able to justify indications for making house calls' load on different factors.

Table 3. Factor loading matrix in domain 3: professional behaviour towards society (n=116)

<table>
<thead>
<tr>
<th>Items</th>
<th>Factors and item factor loadings</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GP trainer Factor 1*</td>
<td>GP trainer Factor 2*</td>
<td>GP trainee Factor 1*</td>
<td>GP trainee Factor 2*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eigenvalue</td>
<td>Eigenvalue</td>
<td>Eigenvalue</td>
<td>Eigenvalue</td>
<td></td>
</tr>
<tr>
<td>The GP trainee:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Bears the consequences of his/her own conduct</td>
<td>0.67</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Is able to justify deviations from rules and guidelines</td>
<td>0.53</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Keeps promises and agreements</td>
<td>0.62</td>
<td>0.57</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Does not hide behind others (give others the blame or responsibility)</td>
<td>0.64</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Is aware of how his/her own norms regarding disease influence disease management</td>
<td>0.51</td>
<td>0.59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Does not impose his/her own norms and values upon others</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Deals meticulously with moral requests for care (e.g., abortion, euthanasia)</td>
<td>-</td>
<td>-</td>
<td>0.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 Is able to set priorities in the choice of topics for quality improvement</td>
<td>0.65</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 Is able to signal suboptimal care within the practice</td>
<td>0.67</td>
<td>0.72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 Is able to work out a quality-improvement project</td>
<td>0.58</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 Is able to estimate which problems are suitable for a quality-improvement project</td>
<td>0.76</td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 Is able to name the tasks to be, or that he/she would like to be, delegated to the assistant</td>
<td>0.64</td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 Has perceptions about how form can be given to means of contact (telephone services, diabetes surgery hours, etc.)</td>
<td>0.76</td>
<td>0.60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 Is able to justify indications for making home visits</td>
<td>-</td>
<td>-</td>
<td>0.49</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>15 Has perceptions about how repeat prescriptions can be written in a responsible manner</td>
<td>-</td>
<td>-</td>
<td>0.49</td>
<td>0.66</td>
<td>0.65</td>
</tr>
<tr>
<td>16 Is aware of the meaning and relative value of scientific evidence in decision-making</td>
<td>0.55</td>
<td>0.48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17 In decision-making, weighs scientific evidence against factors related to the patient or the circumstances</td>
<td>0.67</td>
<td>0.45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 Is able to justify choices made on the basis of scientific evidence</td>
<td>0.63</td>
<td>0.46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 Is able to explain his/her own norms and values regarding the application of scientific evidence</td>
<td>0.52</td>
<td>0.56</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Factors: 1 = responsibility; 2 = quality management
Factor loadings less than 0.40 not shown. Italicized items have been removed from the instrument.

The items 'deals meticulously with moral requests for care (e.g. abortion, euthanasia)' and 'has perceptions about how repeat prescriptions can be written in a responsible way' had greater factor loadings in the trainee group.
The first factor was composed of such behaviours as 'bearing the consequences of his/her own conduct' and 'not hiding behind others' and was labeled responsibility. The content of the second factor included items representing behaviour such as 'being able to signal suboptimal care within the practice' and 'being able to set priorities in the choice of topics for quality management'. This factor was labeled quality management. Two items had a factor loading of just less than 0.40 (table 3) in the trainer group and were removed after reaching consensus. Although the item 'deals meticulously with moral requests for care (e.g. abortion, euthanasia)' had a factor loading less than 0.40 in the trainer group, it was retained because of educational significance and assigned to the first factor responsibility. Tucker's phi values were 0.91 for the factor quality management and 0.94 for responsibility.

**Domain 4: Professional behaviour towards oneself**

The results are shown in table 4. A two-factor solution for both groups was indicated. In the fourth domain, professional behaviour towards oneself, again no major differences in the factor structures between GP trainers and trainees were found.

**Table 4. Factor loading matrix in domain 4: professional behaviour towards oneself (n=116)**

<table>
<thead>
<tr>
<th>Items</th>
<th>Factors and item factor loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GP trainer Factor 1* Eigenvalue</td>
</tr>
<tr>
<td>The GP trainee:</td>
<td>12.07</td>
</tr>
<tr>
<td>1</td>
<td>Is able to name reactions, thoughts, and feelings that patients evoke</td>
</tr>
<tr>
<td>2</td>
<td>Asks questions about his/her own role in relationships (patient, group, gp trainer, etc.)</td>
</tr>
<tr>
<td>3</td>
<td>Uses specific practical situations as starting points for critical self-reflection</td>
</tr>
<tr>
<td>4</td>
<td>In a specific situation with a patient, is able to analyze his/her own behaviour and adjust it if necessary</td>
</tr>
<tr>
<td>5</td>
<td>Dares to express and act upon his/her own point of view</td>
</tr>
<tr>
<td>6</td>
<td>Is able to mention differences of opinion</td>
</tr>
<tr>
<td>7</td>
<td>Discusses his/her shortcomings and failures without losing belief in his/her own competence</td>
</tr>
<tr>
<td>8</td>
<td>Makes a realistic estimation of his/her own strong and weak points</td>
</tr>
<tr>
<td>9</td>
<td>Is able to balance work and private life</td>
</tr>
<tr>
<td>10</td>
<td>Is able to mention aspects of work that increase satisfaction</td>
</tr>
<tr>
<td>11</td>
<td>Is able to seek timely professional help when experiencing personal problems</td>
</tr>
<tr>
<td>12</td>
<td>Is able to cope with feelings of powerlessness in the care process</td>
</tr>
<tr>
<td>13</td>
<td>Is able to give positive as well as negative feedback</td>
</tr>
<tr>
<td>14</td>
<td>Is open about feelings provoked by feedback</td>
</tr>
</tbody>
</table>
### Chapter 2

Factors and item factor loadings

<table>
<thead>
<tr>
<th>Items</th>
<th>GP trainer Factor 1*</th>
<th>GP trainer Factor 2*</th>
<th>GP trainee Factor 1*</th>
<th>GP trainee Factor 2*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Eigenvalue</td>
<td>Eigenvalue</td>
<td>Eigenvalue</td>
<td>Eigenvalue</td>
</tr>
<tr>
<td>15  Attaches importance to what others think about his/her behaviour</td>
<td>0.68</td>
<td>0.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16  Is able to make feedback concrete and specific</td>
<td>0.70</td>
<td>0.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17  Adheres to agreements made during feedback</td>
<td>0.67</td>
<td>0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18  Sets priorities in learning</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19  Does not resist being judged</td>
<td>0.59</td>
<td>0.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20  Has an enquiring mind (asks questions and takes initiative)</td>
<td>0.59</td>
<td>0.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21  Is able to figure things out for him/herself</td>
<td>0.53</td>
<td>-</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>22  Is able to adapt when patients change their minds about a treatment</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>23  Is able to adapt and keep control of the situation if patients unexpectedly need to be seen during other activities</td>
<td>0.51</td>
<td></td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>24  Recovers rapidly after an unpleasant consultation</td>
<td>-</td>
<td>-</td>
<td>0.60</td>
<td>0.64</td>
</tr>
<tr>
<td>25  Is able to admit his/her own mistakes</td>
<td>0.48</td>
<td></td>
<td>0.44</td>
<td>0.47</td>
</tr>
<tr>
<td>26  Takes action to rectify his/her own mistakes</td>
<td>0.52</td>
<td></td>
<td>0.47</td>
<td></td>
</tr>
<tr>
<td>27  Withdraws from the consequences of his/her own mistakes</td>
<td>0.41</td>
<td></td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>28  Is able to cope after making a mistake</td>
<td>-</td>
<td>-</td>
<td>0.66</td>
<td>0.63</td>
</tr>
<tr>
<td>29  Learns from mistakes</td>
<td>0.51</td>
<td>-</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>30  Is able to let a mild disorder (e.g., tiredness) run its own course even though the correct diagnosis is a mystery</td>
<td>0.58</td>
<td></td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>31  Makes rational deliberations about whether it is necessary to request specialist or other advice</td>
<td>0.62</td>
<td>-</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>32  Is able to deal with the possibility that a treatment decision may be unsuccessful</td>
<td>0.63</td>
<td>-</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>33  Expresses strong emotions by means of words instead of actions</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>34  Is able to deal with difficult or angry patients</td>
<td>0.59</td>
<td></td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>35  Is able to conduct interventions that lead to a decrease in aggression from the patient</td>
<td>0.52</td>
<td>-</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>36  Is able to formulate his/her own opinion in a clear and inoffensive manner</td>
<td>0.61</td>
<td>-</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>37  Keeps an eye on his/her own safety adequately</td>
<td>-</td>
<td>-</td>
<td>0.52</td>
<td></td>
</tr>
</tbody>
</table>

*Factors: 1 = reflection and learning and 2 = dealing with emotions

Factor loadings less than 0.40 not shown. Italicized items have been removed from the instrument.

The items 'discusses one's shortcomings and failures without losing belief in one's own competence', 'makes a realistic estimation of one's own strong and weak points', 'is able to mention aspects of work that increase satisfaction', 'is able to cope with feelings of powerlessness in the care process', and 'learns from one's own mistakes' load on different factors.
The second factor in the GP trainer group has approximately the same items as the factor of the GP trainees. The only exception is the item 'learns from one's mistakes'.

The first factor included items that reflect behaviours such as 'being able to name thoughts and feelings that patients evoked in oneself', 'being able to analyze one's own behaviour in specific situations' and 'being able to figure things out by oneself' and was labeled reflection and learning. The second factor consisted of items such as 'being able to cope after making a mistake' and 'being able to deal with the possibility that a treatment is unsuccessful'. This factor was labeled dealing with emotions. After discussion, we decided to assign the items 'discusses one's shortcomings and failures without losing belief in one's own competence', 'makes a realistic estimation of one's strong and weak points', 'is able to mention aspects of work that increase satisfaction' and 'learns from one's own mistakes' to the scale reflection and learning. The item 'is able to cope with feelings of powerlessness in the care process' was assigned to the scale dealing with emotions. These decisions were based on face validity.

Six items had factor loadings less than 0.40 (table 4). The items 'set priorities in learning' and 'be able to balance work and private life' have an important educational significance, so they were retained and assigned to the factor reflection and learning. The other items were removed from the list.

Tucker’s phi values were computed for both factors; they were 0.90 and 0.91.

**Internal consistency**
Table 5 shows the Cronbach’s α coefficients for the two groups of participants in our study (GP trainers and GP trainees).

The Cronbach’s α coefficients for the GP trainer sample ranged from 0.79 (dealing with emotions) to 0.95 (reflection and learning), which indicates good to excellent internal consistency within each factor. Good to excellent internal consistency was also found in the trainee group, with values ranging from 0.72 (professional distance) to 0.91 (reflection and learning).

**Table 5. Cronbach’s α associated with each factor (n=116)**

<table>
<thead>
<tr>
<th>Domain: professional behaviour towards</th>
<th>Factor</th>
<th>Number of items</th>
<th>Cronbach’s α GP trainers</th>
<th>Cronbach’s α GP trainees</th>
</tr>
</thead>
<tbody>
<tr>
<td>The patient</td>
<td>Respecting patient’s interests</td>
<td>20</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td>The patient</td>
<td>Professional distance</td>
<td>5</td>
<td>0.82</td>
<td>0.72</td>
</tr>
<tr>
<td>Other professionals</td>
<td>Collaboration skills</td>
<td>10</td>
<td>0.91</td>
<td>0.86</td>
</tr>
<tr>
<td>Other professionals</td>
<td>Management skills</td>
<td>9</td>
<td>0.94</td>
<td>0.87</td>
</tr>
<tr>
<td>The society</td>
<td>Responsibility</td>
<td>10</td>
<td>0.78</td>
<td>0.82</td>
</tr>
<tr>
<td>The society</td>
<td>Quality management</td>
<td>7</td>
<td>0.89</td>
<td>0.87</td>
</tr>
<tr>
<td>Oneself</td>
<td>Reflection and learning</td>
<td>23</td>
<td>0.95</td>
<td>0.91</td>
</tr>
<tr>
<td>Oneself</td>
<td>Dealing with emotions</td>
<td>9</td>
<td>0.79</td>
<td>0.86</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td></td>
<td><strong>93</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 2

Discussion

The results of this study provide psychometric support for the Nijmegen Professionalism Scale. Previous results\textsuperscript{12} supported the content validity of the instrument as well as its feasibility as a tool to educate for professionalism.

The original structure, based on consensus and face validity alone, was not replicated\textsuperscript{12}. Instead, a much simpler structure, with two scales for each domain, was found. In our view, this new structure makes the instrument conceptually clearer and more straightforward for GP trainers and GP trainees to use.

In contrast with the traditional approach, competency-based medical education can potentially lead to individualized flexible training, transparent standards and increased public accountability. If applied inappropriately, it can also result in demotivation, focus on minimally acceptable standards, increased administrative burden and a reduction in the educational content. Higher-order competencies, such as professionalism, need to be defined and developed more robustly\textsuperscript{4}. Professional behaviour is a complex construct to define, and without consensus of this construct, teaching and assessing professional behaviour are problematic. We compared the factor structure of self-assessment and external evaluations. No considerable differences in the four domains were found. This indicates that GP trainers and trainees attach the same meaning to the construct of professional behaviour, which creates a solid foundation for effective teaching and assessing of this essential part of medical performance.

Feasibility is the most common limitation, since assessment tools often take a lot of time. Some concern remains about the feasibility of the Nijmegen Professionalism Scale; the final list contains 93 items and may be too time consuming for most GP trainers. However, users were asked to comment on the instrument\textsuperscript{12}. They appreciated the valuable input it provided during the tutorial. No one criticized the length of the list. The Nijmegen Professionalism Scale is designed to guide professional growth. The specific details of the instrument enable trainers not only to assess, but also to encourage and monitor specific behaviour. As professional behaviour should be observed from the beginning of the training, feedback in an early stage of the training allows GP trainees to remedy possible lack of professional behaviour. Nevertheless, exploring whether the number of items could be reduced would be worthwhile. In addition, as the Nijmegen Professionalism Scale consists of four domains, it is possible for users to administer one domain at the time.

GP trainer and trainee work in close cooperation for an extended period of time and this may undermine the independence of the scores. Although it is relatively time consuming to complete the instrument, the use of instruments like the Nijmegen Professionalism Scale have to be considered as a very important element in the development of their trainees. GP trainers must
be aware that only by using the Nijmegen Professionalism Scale in the appropriate way, they formatively support the development of professional behaviour of their trainees in a integrated, coherent and longitudinal fashion. This issue underscores the importance of rater training in the accurate use of assessment instruments before implementing them.

This study to validate the Nijmegen Professionalism Scale represents one further step. Van de Camp et al. already tested the content validity of the instrument in a qualitative study\(^{12}\). However, to construct a sound validity argument, a much broader range of evidence is required. It has been argued that we cannot infer validity from a single analysis\(^{23}\). Further information supporting the validity of the Nijmegen Professionalism Scale, for instance, would be data suggesting that it accurately identifies trainees with performance deficits and that the instrument is able to measure the professional growth of the trainees. Since we did not detect ceiling effects growth can in principle be assessed, but we did not test this.

**Conclusion**

Meaningful, reliable and valid assessment is crucial in the promotion of professionalism in GP trainees. On the basis of this study, we can conclude that the Nijmegen Professionalism Scale is a reliable tool to assess their professional behaviour. The results of this study show that GP trainers and trainees agree on the definition and meaning of professional behaviour. We consider the Nijmegen Professionalism Scale to be a promising tool for assessing and enhancing the professional behaviour of GP trainees.
References

Chapter 3

Assessment of CanMEDS roles in postgraduate training: The validation of the Compass

F. Tromp
M.J.F.J. Vernooij-Dassen
R.P.T.M. Grol
A.W.M. Kramer
B.J.A.M. Bottema

Patient Education and Counseling 2012;89:199-204.
Abstract

Objective: In medical education the focus has shifted from gaining knowledge to developing competencies. To effectively monitor performance in practice throughout the entire training, a new approach of assessment is needed. This study aimed to evaluate an instrument that monitors the development of competencies during postgraduate training in the setting of training of general practice: the Competency Assessment List (Compass).

Methods: The distribution of scores, reliability, validity, responsiveness and feasibility of the Compass were evaluated.

Results: Scores of the Compass ranged from 1 to 9 on a 10-point scale, showing excellent internal consistency ranging from .89 to .94. Most trainees showed improving ratings during training. Medium to large effect sizes (.31–1.41) were demonstrated when we compared mean scores of three consecutive periods. Content validity of the Compass was supported by the results of a qualitative study using the RAND modified Delphi Method. The feasibility of the Compass was demonstrated.

Conclusion: The Compass is a competency based instrument that shows in a reliable and valid way trainees' progress towards the standard of performance.
Introduction

In medical education the focus has shifted from the acquisition of knowledge to the development of competencies\(^1\). The rationale behind this educational shift was that competency-based education should better prepare medical doctors for clinical practice\(^2\). Competency-based teaching has been driven by competency frameworks such as the Accreditation Council for Graduate Medical Education and the American Board of Medical Specialties (ACGME/ABMS) competencies\(^3\), and the Canadian Medical Education Directives for Specialists (CanMEDS) 2000, issued by Canada’s Royal College of Physicians and Surgeons (Royal College of Physicians and Surgeons of Canada 2000)\(^4\).

Competencies have been described as the "habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in daily practice for the benefit of the individual and the community being served"\(^5\). This definition emphasizes the operational nature of competencies, they are not static knowledge, memorization of facts, but rather the behaviours and nature of actions that are carried out while utilizing medical knowledge. Competency is both what one knows and how one acts when exercising that knowledge. Moreover, competencies have to be demonstrable and involve performing tasks that can be witnessed by evaluators and educators.

With the transformation into a competency-based program, it is equally important that parallel strategies are chosen to assess these competencies. In 1990, George Miller proposed a classification, in the form of a pyramid, for methods of assessment in medical education. 'Knows' is at the base of the pyramid followed by 'knows how', 'shows how', and 'does'\(^6\). The ultimate goal for a valid assessment of competencies is to test how the trainee actually performs in practice, represented by the performance or "does level", the top of the pyramid.

In daily practice, assessing competencies by breaking down the competency into smaller units which are assessed separately should be avoided\(^7\). Most current assessment tools, however, measure only a part of a competency, such as for instance medical knowledge.

No single instrument can assess all competencies. Moreover, one single assessment has limitations, such as case-specificity, or low reliability\(^8\). This insight inspired Van der Vleuten et al. to advocate programmes of assessment\(^7\). In a program of assessment multiple sources of information from various methods are used to construct an overall judgment by triangulating information across these sources. These assessments should also generate feedback, because feedback promotes learning: it informs trainees of their progress or lack thereof; it advises trainees regarding observed learning needs and resources available to facilitate their learning; and it motivates trainees to engage in appropriate learning activities\(^9\).
In this context the Competency Assessment List ('Compass') was developed. The Compass aggregates assessments of performance in practice during a postgraduate training programme. It follows a programmatic approach by aggregating separate assessments of performance in practice\(^7,8\): in order to continuously gauge their trainees' competencies, trainers are to use multiple assessment instruments, on several occasions, in various working situations throughout the entire training. Every three months, the aggregated results of these assessments are summarized in the Compass and are discussed between trainer and trainee. These progress meetings should promote habits of learning and self-reflection. With the use of the Compass, continuous assessment becomes a key element in the training.

The aim of this study was to explore the reliability, validity, and feasibility of the Compass. We examined the psychometric properties of the Compass and evaluated if the Compass discriminates trainees' strengths and weaknesses, and progress towards the required professional standard of performance.

**Methods**

**Context of the study**

We performed our study in Dutch Postgraduate training for general practice. This training has a nationally endorsed curriculum. The Departments of General Practice of the eight faculties of medicine are responsible for the organization of a 3-year postgraduate training. The first and the last year consist of internships in general practice, the second year in hospitals and other medical institutions. During the whole training period, one day per week is for day-release courses and self-directed learning. This study has been conducted during the first year of training.

**The Compass**

The Compass was designed to identify trainees' performance throughout the training, as well as the trainees' progress towards the standard of performance required upon completion of the training, thus providing formative assessment: guiding future learning, encouragement, and promoting reflection. The seven CanMEDS roles are assessed in an integrated, coherent and prospective fashion.

Every three months ratings are given on a 10 point scale using a fixed reference point for judgments. This reference point is the standard expected by the trainer upon final completion of the three-year training. The awarded scores are based on an aggregated assessment of the preceding three-month period and represent the degree of competence mastered by the trainee. A question mark can be noted if there is too little relevant information available about the trainees' performance indicating that the assessor has not made enough observations or was not able to gather enough information for assessment.
Each of the seven roles are subdivided into competencies. Box 1 gives an overview. The competencies are accompanied by a list of examples of appropriate behaviours indicating mastering of the competence under consideration.

As a general quality requirement, the trainer is expected to gather information about the performance of the trainee:

- at various moments throughout the designated three months training period,
- in different working situations, such as patient consultations, home visits, the provision of services, telephone conversations with patients and specialists and practice meetings,
- from different sources, such as practice staff, patients, self reflection reports and referral letters,
- with the aid of various observation based assessment instruments, such as the mini-CEX\textsuperscript{10}, Maas Global\textsuperscript{11}, the Nijmegen Professionalism Scale\textsuperscript{12} and by direct observation.

Box 1. Overview of the seven roles. The roles are subdivided into several competencies, accompanied by a list of examples of appropriate behaviours

<table>
<thead>
<tr>
<th>Competencies</th>
<th>Examples of behavioural indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Medical Expertise</strong></td>
<td></td>
</tr>
<tr>
<td>Interprets symptoms in context</td>
<td>Mentions contextual factors noted in the medical file that may be related to the complaint if necessary. Inquires about the physical and psychological background to the complaint if necessary. Takes the patient's age and medical history into account</td>
</tr>
<tr>
<td>Applies the diagnostic, therapeutic and preventive arsenal of the profession in an appropriate and evidence-based way</td>
<td>Establishes whether the nature of the problem does or does not merit a general practice intervention. Makes correct (provisional) diagnoses (ICPC A to Z) based on his or her findings, physical examination and additional diagnostic procedures.</td>
</tr>
<tr>
<td>Provides primary care in a systematic way</td>
<td>Has mastered the complete spectrum of identifying the problem, taking a history, carrying out physical examinations and other tests, providing information and advice, supervision and referral. Provides care in an appropriate way during consultations and home visits by adopting a logical approach, first gathering information, then making a (provisional) diagnosis and deciding on treatment or arranging for treatment to be carried out.</td>
</tr>
<tr>
<td><strong>Communication</strong></td>
<td></td>
</tr>
<tr>
<td>Develops effective treatment relationships with patients</td>
<td>Allows the patient to give their own story or account of the situation. Shows dedication through his or her body language, eye contact and tone of voice.</td>
</tr>
<tr>
<td>Applies communication techniques and resources appropriately</td>
<td>Is the manager of the conversation with the patient (clear agenda, logical structure, summarizing, keeping sight of the time) Checks if patient understood the provided information.</td>
</tr>
<tr>
<td>Ensures that the patient is actively involved in the decision-making</td>
<td>Exploration Clarifies the patient's request for help/care. Agreement regarding the approach Agrees the treatment proposal or approach with the patient.</td>
</tr>
<tr>
<td>Collaboration</td>
<td>Management</td>
</tr>
<tr>
<td>------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Contributes to effective intra- and interdisciplinary collaboration</td>
<td>Provides integral and appropriate general practice care that is continuous and accessible</td>
</tr>
<tr>
<td>Applies collaboration skills appropriately</td>
<td>Applies organizational and management principles appropriately</td>
</tr>
<tr>
<td>Makes appropriate referrals on the basis of a current insight into the expertise of other care providers</td>
<td>Uses information technology for optimal patient care</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sample and procedure
As a part of the training, GP trainers have to complete the Compass every three months. All 76 GP trainers of the first year and their GP trainees associated with the Department of Postgraduate Training for General Practice of the Radboud University Nijmegen Medical Centre were asked to participate by handing over copies of completed Compass lists to the researchers. Trainers and trainees were informed about the aim of the study by one of the researchers (FT). Participants could withdraw from the study at any time without any consequence. Participants consented to provide their anonymous data for research analysis. Since this study did not involve patients no ethical approval had to be requested. For this study, we used data of the Compass completed by GP trainers in 2010 of three consecutive 3-month periods in the first year of the training The first three-month periods is named T1, the second period T2 and the third three-month period T3.

Distribution of scores
The distribution of scores provides information about the discriminative power of an item as well as the occurrence of floor and ceiling effects\textsuperscript{13}. We considered floor effects to occur if trainees' scores are repeatedly on the lowest end of the scale (score 1) in T1 and in T2 as well as in T3. Ceiling effects were defined as scores on the highest end of the scale (score 10) in T1 and T2 as well as in T3.

Reliability
Cronbach's alpha was computed to determine the internal consistency of items within each of the seven roles. We considered a Chronbach's alpha of at least .70 an indication of satisfactory reliability of each scale\textsuperscript{14}. To examine the homogeneity of the scales we computed the item-total scale correlation. Item-total correlations of .40 or higher was considered acceptable evidence of contribution of each item to the scale homogeneity.

Content validity
The content validity of the Compass was tested in a qualitative study using the RAND modified Delphi Method\textsuperscript{15}. First, we selected an 8-member expert panel. For practical reasons, we limited the number to eight. Members of the panel should have experience with the use of the Compass. Therefore, we approached seven of the eight Departments of General Practice. The selected panelists were 6 general practitioners trainers and 2 GP-trainees. The first version of the Compass was developed in advance by a consensus finding process among assessment experts of the institutes for general practice training, using the CANMEDS guidelines. This version was sent to the panelists as preliminary instrument (the preliminary instrument had been used several years, but the content has not been validated so far). In a two round Modified Delphi process, we asked the panelists to judge the extent to which the behavioural indicators were suitable examples of appropriate behaviour in order to assess the role of the GP-trainee. In the first
round, the panel members individually rated the behavioural indicators of each CanMEDS role on a nine-point scale, ranging from extremely unsuitable (=1) to extremely suitable (=9). In addition to rating the behavioural indicators, panelists were asked to comment on the behavioural indicators and suggest revisions. Roles and competencies were not assessed by the panel because legislation requires that all postgraduate medical training programs in the Netherlands are based on these roles and competencies. In the second round, the panel members met under the leadership of a moderator. During the meeting, the panelists discussed their previous ratings, focusing on areas of disagreement, and were given the opportunity to modify the original list. Disagreement among the panelists was defined if at least one third of the panel members rated behavioural indicators in the range 1–3, while at least one-third of the others rated the behavioural indicators in the range 7–9, according to accepted criteria in the RAND modified Delphi procedure. After discussing each one of the seven roles, they rerated each behavioural indicator individually. No attempt was made to force the panel to consensus. Behavioural indicators were accepted when they received a mean score of 7 or higher.

**Responsiveness**

Responsiveness is defined as 'the ability of an instrument to detect change over time in the construct to be measured'\(^1\)\(^3\). During training, trainees' performance is expected to improve. This improved performance should be reflected in higher scores on the Compass. The ability of the Compass to identify trainees' progress towards the standard of performance required upon completion of the training was evaluated by comparing the average Compass scores at the end of each three-month interval. The increase in ratings throughout the year represents the progress made during training. We used a one-way repeated measures ANOVA to determine if the ratings of three successive periods differed significantly from each other. To examine the pattern of change, Bonferroni post hoc tests were executed to establish if ratings of the first three months period (T1) differed significantly from ratings of the second period (T2), and if ratings from the second three months period differed from ratings of the third period (T3). In addition, we calculated effect sizes. Effect size (ES) is an index that quantifies the degree to which the results should be considered negligible or important\(^1\)\(^6\). ES values are classified in three categories: small (ES ≈ .20), medium (ES ≈ .50) and large (ES ≈ .80)\(^1\)\(^6\).

**Feasibility**

Feasibility was evaluated by examining how many question marks were noted for each competency.

**Results**

**Sample**

Sixty-eight out of seventy-four GP trainees provided lists of three consecutive three-month periods. This resulted in 184 lists that were eligible for inclusion.
Missing data (20 lists) were due to absence of trainees because of illness (2 trainees), or pregnancy (4 trainees).

The distribution of scores

Table 1 represents the descriptive statistics for the scores of the Compass. All roles showed a relatively large standard deviation compared to mean scores. Scores ranged from 1 to 8 except scores on Professionalism, which ranged from 1 to 9. Excellent performance (scores on the high end of the scale) as well as poor performance (scores on the low end of the scale) could be identified.

In the first three-month period 1.6–4.7% of the trainees scored at the lowest end of the scale. In the subsequent periods, however, no scores at the lowest end were found. So, no floor effects were found. We did not detect ceiling effects either.

Table 1. Descriptive statistics for the scores of the Compass. Feasibility: percentage of question marks after each three month period

<table>
<thead>
<tr>
<th>Role</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Expertise</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean scores</td>
<td>4.43</td>
<td>5.43</td>
<td>6.18</td>
</tr>
<tr>
<td>SD</td>
<td>1.51</td>
<td>1.17</td>
<td>1.04</td>
</tr>
<tr>
<td>Range</td>
<td>1-8</td>
<td>3-8</td>
<td>4-8</td>
</tr>
<tr>
<td>% at Floor</td>
<td>1.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% at ceiling</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Question marks</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Communication</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean scores</td>
<td>4.53</td>
<td>5.44</td>
<td>6.02</td>
</tr>
<tr>
<td>SD</td>
<td>1.70</td>
<td>1.21</td>
<td>1.25</td>
</tr>
<tr>
<td>Range</td>
<td>1-8</td>
<td>3-8</td>
<td>3-8</td>
</tr>
<tr>
<td>% at Floor</td>
<td>3.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% at ceiling</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Question marks</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Management</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean scores</td>
<td>4.51</td>
<td>5.41</td>
<td>6.07</td>
</tr>
<tr>
<td>SD</td>
<td>1.92</td>
<td>1.43</td>
<td>1.18</td>
</tr>
<tr>
<td>Range</td>
<td>1-8</td>
<td>2-8</td>
<td>3-8</td>
</tr>
<tr>
<td>% at Floor</td>
<td>1.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% at ceiling</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Question marks</td>
<td>9.4%</td>
<td>2.3%</td>
<td>4.5%</td>
</tr>
<tr>
<td>Collaboration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean scores</td>
<td>4.41</td>
<td>5.43</td>
<td>6.11</td>
</tr>
<tr>
<td>SD</td>
<td>1.88</td>
<td>1.40</td>
<td>1.31</td>
</tr>
<tr>
<td>Range</td>
<td>1-8</td>
<td>3-8</td>
<td>3-9</td>
</tr>
<tr>
<td>% at Floor</td>
<td>1.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% at ceiling</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Question marks</td>
<td>7.8%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>
### Social Accountability

<table>
<thead>
<tr>
<th>Role</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean scores</td>
<td>4.07</td>
<td>4.87</td>
<td>5.49</td>
</tr>
<tr>
<td>SD</td>
<td>1.85</td>
<td>1.43</td>
<td>1.50</td>
</tr>
<tr>
<td>Range percentage</td>
<td>1-8</td>
<td>2-7</td>
<td>2-8</td>
</tr>
<tr>
<td>% at Floor</td>
<td>4.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% at ceiling</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Question marks</td>
<td>15.6%</td>
<td>9.1%</td>
<td>2.3%</td>
</tr>
</tbody>
</table>

### Science and Education

<table>
<thead>
<tr>
<th>Role</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean scores</td>
<td>4.36</td>
<td>5.15</td>
<td>5.65</td>
</tr>
<tr>
<td>SD</td>
<td>1.99</td>
<td>1.81</td>
<td>1.63</td>
</tr>
<tr>
<td>Range percentage</td>
<td>1-8</td>
<td>2-8</td>
<td>2-8</td>
</tr>
<tr>
<td>% at Floor</td>
<td>4.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% at ceiling</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Question marks</td>
<td>31.2%</td>
<td>13.6%</td>
<td>9.1%</td>
</tr>
</tbody>
</table>

### Professionalism

<table>
<thead>
<tr>
<th>Role</th>
<th>Cronbach’s α</th>
<th>Range of item-total correlation</th>
<th>F-values</th>
<th>Effect Size T2-T1</th>
<th>Effect size T3-T1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical expertise</td>
<td>.92</td>
<td>.78-.90</td>
<td>F(1.599, 73.55) = 111.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communication</td>
<td>.91</td>
<td>.72-.94</td>
<td>F(1.581, 120.24) = 265.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Management</td>
<td>.89</td>
<td>.72-.91</td>
<td>F(1.50, 151.54) = 122.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collaboration</td>
<td>.91</td>
<td>.73-.93</td>
<td>F(1.30, 137.87) = 201.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social accountability</td>
<td>.93</td>
<td>.82-.94</td>
<td>F(1.38, 139.58) = 161.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Science and education</td>
<td>.94</td>
<td>.90-.96</td>
<td>F(1.40, 122.08) = 111.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professionalism</td>
<td>.91</td>
<td>.77-.93</td>
<td>F(1.49, 169.34) = 173.87</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

T1= first three months period, T2= months 4, 5 and 6, T3= months 7,8 and 9.

**Reliability**

Cronbach’s alpha and homogeneity scales are shown in table 2. All scales showed excellent internal consistency ranging from .89 to .94. All items correlated sufficiently with each other to form a scale.

**Table 2.  Results of reliability analysis and range of item-total correlation of all roles of the Compass, F and P values of the repeated measures ANOVA and Effect Sizes T2-T1 and T3-T1**
Content validity
The panel reached a consensus on 46 of the 80 behavioural indicators in the first, written rating round. Thirty-four behavioural indicators were discussed in the second round. More specifically, 5 behavioural indicators of the competence Medical Expertise, 3 behavioural indicators of Communication, 5 of Collaboration, 4 of Management, 5 of Social Accountability, 4 of Science and Education, and 8 behavioural indicators of Professionalism.

The panel rephrased 10 behavioural indicators in order to improve precision and clearness: 1 of the competence Medical Expertise, 1 of the competence Communication, 2 of the competence Collaboration, 1 of the competence Management, 3 of the competence Social Accountability, 1 of the competence Science and Education and 1 of the competence Professionalism. Also, 5 behavioural indicators were added: 1 at the competence Communication, 2 at Management, 1 at the competence Social Accountability and 1 at Professionalism.

The panel re-rated the behavioural indicators at the end of this meeting, resulting in the removal of 5 indicators: 3 from Medical Expertise, 1 from Collaboration and 1 from Professionalism. The final list contained 80 behavioural indicators.

Responsiveness
The results show that mean scores of all roles in T1, T2 and T3 differed significantly. F and p-values are presented in table 2. Post hoc testing revealed that mean ratings of all roles in T2 were significantly higher than ratings in T1 and, in addition, that mean ratings in T3 were significantly higher than ratings in T2.

Compared with T1, most trainees received higher ratings in T2. A relatively small group of trainees with no increase in ratings was found in all roles: Medical Expertise (21.3%), Communication (31.3%), Management (22.5%), Collaboration (23.8%), Social Accountability (22.5%), Science and Education (22.5%) and Professionalism (16.3%). We also encountered a small amount of negative outcomes reflecting declining performance. These declining ratings were present in Communication (2.5%), Management (1.3%), Collaboration (2.5%), Science and Education (5.0%) and Professionalism (1.3%).

Compared with T2, we also found higher ratings in T3. Trainees with no increase in ratings were found in Medical Expertise (28.8%), Communication (32.5%), Management (27.5%), Collaboration (22.5%), Social Accountability (26.3%), Science and Education (27.5%) and Professionalism (17.5%). Declining ratings were present in the competencies Communication (2.5%) and Management (1.3%).
The differences were not only significant, also the effect sizes were medium to large comparing T1 and T2. The largest ES (.82) was found in Medical Expertise, whereas the smallest ES (.31) was found in Science and Education. Effect sizes were all large when we compared T1 and T3 (Table 2). The largest ES (1.41) was also found in Medical Expertise, and again the smallest ES (.70) was found in Science and Education.

Is the Compass feasible in practice?
To establish feasibility the amount of question marks was explored. Table 1 shows the results. At T1 the most question marks (indicating that the trainer was not able to assess this competency) were found in Management, Collaboration, Science and Education and Social Accountability. During the next periods trainers were able to assess all competencies. Science and Education, however, is after nine months not yet assessed in almost ten percent of the cases.

Discussion and conclusion
Discussion
Our results support the reliability, validity and feasibility of the Compass. There are no indications that ceiling or floor effects occur. Also, the full range of the scale is used. Therefore, the Compass seems suitable to measure trainees’ progress and to identify excellent as well as poor performance. The examination of the content validity resulted in only minor changes of the content of the Compass. This comes as no surprise as the Compass was previously developed by a working group that was composed of members who all were highly experienced in the field of education and assessment.

This is the first instrument, assessing the does level of Miller’s pyramid6, that successfully demonstrates progress of performance of all competencies of GP-trainees. Trainees’ scores are significantly higher each following period. Moreover, the differences are meaningful and relevant.

When progress can be visualized as ratings increase with time, it may have a positive effect on the motivation of the individual towards improving his or her own performance17. Since the Compass provides visible and tangible information of how GP-trainees are progressing, it is suitable for monitoring, discussing and assessing trainees’ performance during training. It provides a clear insight where progress is needed and where the required standard is reached.

The results of this study indicate that by using a fixed-reference point for trainers’ judgments, the Compass makes it possible to visualize progress over time. Prescott-Clements and colleagues argued that instruments which do not use the end of the training as a reference, show little difference in scores during training17,18. There are very little studies that demonstrate progress of performance in daily practice during medical training. Kramer et al. for instance, did not succeed
to demonstrate growth of trainees' communication skills during postgraduate training in general practice, although trainees felt more confident in performing communication skills at the end of the training than at the start of it\textsuperscript{19}.

Often, throughout medical training individual assessment instruments for separate parts of competencies are used\textsuperscript{7}. The results of these assessments may be considered as snapshots in time. It has been established that braking up the competencies into isolated steps, is not a very sensitive method for detecting differences in expertise\textsuperscript{8}. The Compass covers assessments made in a period of three months. This continuous way of assessment proves to be sensitive to change.

A possible drawback of continuous assessments with rating scales is that raters will anticipate that trainees' scores will increase with time and this will positively bias ratings awarded later in the training\textsuperscript{17}. For that reason it was interesting to see that a decrease in mean ratings could be observed in all CanMEDS roles in a number of trainees.

The feasibility of the Compass was demonstrated by the low number of question marks. Trainers were able to assess all CanMEDS roles. In the first three months of training some missing ratings were found in Management, Collaboration, Social Accountability and Science and Education. During the next periods all roles were assessed, although the assessment of Science and Education remained somewhat problematic. This is not unexpected because Science and Education is not a main subject in the first year of the training.

**Conclusion**

In conclusion, the Compass shows trainees' progress towards the standard of performance required upon completion of the training and it is able to identify excellent as well as poor performance. It provides a clear insight where progress is needed and where the required standard is reached. The Compass, a programmatic approach of continuous evaluation, enables GP trainers to assess all the competencies of the GP trainee in a reliable, valid and feasible way.

**Limitations of the study**

A limitation of this study is that our sample is from one of the eight institutions in the Netherlands and therefore may not be fully representative of the whole population of GP trainers and trainees. Further research can provide insight in the generalizability of our results. Extending measurements to all institutions as well as to other postgraduate trainings would enhance the interpretation of transferability of our results. However, uniformity in scale and the same point of reference are required to do so.

In assessing the feasibility of the instrument we counted the question marks. Because feasibility is an important requirement for the implementation and use of instruments, this subject could
do with more information about, for instance, the reason why trainers could not complete the list. To obtain additional information further research is needed. An appropriate method would be to interviewing the users of the Compass.

**Practice implications**  
The Compass is able to monitor progress of competencies during training and to provide feedback to trainees. The programmatic approach of the Compass could be applied in postgraduate training of other specialties provided that the instrument is tailored to the specific needs of that field.
References

Chapter 4

How do trainers form an overall judgment of trainees performance and progress

F. Tromp  
D.A.C.M. van Roermund  
M.J.F.J. Vernooij-Dassen  
B.J.A.M. Bottema  
A.W.M. Kramer
Abstract

Introduction: Comprehensive competency based medical education necessitates a robust and multifaceted assessment system. An instrument was developed that monitors the development of competencies during postgraduate training in the setting of general practice training: the Competency Assessment List (Compass). Trainers are to use multiple assessment instruments on several occasions and aggregate the results to come to an overall judgment. The aim of this study was to gain more insight into how trainers form their judgment of trainees performance and progress.

Methods: We performed a qualitative content analysis of the written considerations about judgments using a constant comparative method approach. To investigate differences between the competency domains, we did not only consider the emerging themes but also the number of codes as a measure of elaborateness.

Results: The coding of the written comments resulted in 5 themes (Judgment; Provisional judgment; Judgment including other competencies; Giving directions; and Impressions) from which two categories of competencies emerged: Medical Expertise, Communication and Professionalism (first category) and Collaboration, Management, Social Accountability and Science & Education (second category). Considerations of the first category appeared to be more specific, focused on performance, development and personal characteristics. In the second category provisional judgments were given. In addition, judgments of these competencies seemed to be blurred, as other competencies were included in the considerations.

Discussion and conclusion: Two categories of competencies could be distinguished. The written content of the perhaps more familiar competencies was more extensive and elaborate than the content of the 'new' ones. We have speculated on a number of reasons why. Further study should result in finding appropriate ways to assess the competencies Collaboration, Management, Social Accountability and Science & Education.
Introduction
Competency frameworks such as the one developed by the Accreditation Council for Graduate Medical Education and the American Board of Medical Specialties (ACGME/ABMS)\(^1\), and the Canadian Medical Education Directives for Specialists (CanMEDS) 2000\(^2\), guide the construction of the curriculum in many countries today. Stimulated by the widespread implementation of competency-based curricula, parallel strategies are chosen to assess these competencies\(^3\).

Competency-based medical education necessitates a robust and multifaceted assessment system\(^4\)-\(^7\). For just over two decades leading educationists, including medical educators, have highlighted the intimate relationship between learning and assessment\(^5\). Indeed, in an educational context it is now argued that learning is the key purpose of assessment\(^6\). Evaluations or judgments that arise from it are important for the trainees. Moreover, in competency-based education learning is supposed to be self-directed. From the literature we know that it is hard to get people to self-direct\(^8\)-\(^10\). Also, evidence now suggests that accurate self-assessment may be an unrealistic goal\(^11\). Several reviews\(^12\),\(^13\), conclude that self-assessment undertaken as an individually conducted internal activity has little accuracy or reliability. Additionally, those with the least proficiency in a domain tend to overestimate their capacities most and seem to be inaccurate self-assessors\(^14\). Eva and Regehr\(^13\) reviewed self-assessment from various theoretical perspectives, one of which was social cognitive theory, which acknowledges that much of human learning occurs within a social environment\(^15\). Learning about ourselves requires looking outward to that environment, especially to the responses of others to our behaviour, and using this feedback to inform our assessments of ourselves. For that reason, in the supervision of trainees to support self-directed learning, feedback from trainers plays an important role. This feedback can be used to plan new learning tasks or goals based on insufficient performance\(^16\). So, competency-based medical education requires enhanced attention to assessment to ensure that trainees receive frequent and high-quality feedback to guide their progress and the acquisition of the necessary competencies\(^17\). This also implicates that feedback resulting from assessment should not only be quantitative, but also qualitative, descriptive and narrative to inform trainees with detailed and specific information about performance.

In the context of postgraduate medical training nowadays, assessment follows increasingly a programmatic approach\(^6\). In a programmatic approach assessment requires a deliberate and arranged set of longitudinal assessment activities. So, various instruments, sources and information are used and aggregated to come to an overall judgment. Although single-point assessments are useful in generating valuable feedback and support, they do not give a clear picture of professional development of the trainee\(^3\),\(^6\),\(^18\).

In the setting of general practice training in the Netherlands The Competency Assessment List (Compass) was developed. It follows a programmatic approach: trainers are to use multiple
Chapter 4

assessment instruments on several occasions and aggregate the results in order to come to an overall judgment. The main purpose of the Compass is to support learning (assessment for learning), by providing trainees constructive feedback. This feedback is discussed every three months in progress meetings, and should promote habits of self-reflection and lifelong learning in order to better prepare trainees for clinical practice19. With the use of the Compass, continuous assessment becomes an important element in the curriculum. The Compass is described more elaborately in the methods section. The Compass was introduced in the Netherlands in all departments of general practice-training in 2007. It was developed by a consensus finding process among experienced GP’s and medical educators of the general practice training from the eight departments for general practice training in the Netherlands. Each department was responsible for the implementation and use of the instrument.

In an earlier validation study, the distribution of scores, reliability, validity, responsiveness and feasibility of the Compass were evaluated20. In this paper, we want to gain more insight into how trainers form their judgment of trainees performance and progress by examining the considerations they have written on the assessment form. These considerations are written down to make clear what trainers have seen that has led to a quantitative judgment. These written comments shape the feedback provided to the trainee. We studied the content of the written comments and examined if the content of the comments differed between competency domains.

The results of this study may contribute to maximize the quality of assessment in competency-based medical education. In addition the results of this study may improve the design of assessment instruments.

Methods

Context of the study

Postgraduate training for general practice in the Netherlands has a uniform structure and a nationally endorsed curriculum. The departments of general practice of the eight faculties of medicine in the Netherlands, in the context of postgraduate training, usually addressed as training institutes, are responsible for the organization of the training. To become a GP in the Netherlands, 3 years of training are required after graduation as a medical doctor. The first and the last year are spent in general practice, the second year in hospitals and other medical institutions. During the whole training period one day a week is reserved for day-release courses to support learning in practice. This study focuses on the outcomes of the first year.

Sample and procedure

GP trainers are required to complete the Compass to evaluate their trainee every three months. All 76 GP trainers of the first year and their GP trainees associated with the Department of
Postgraduate Training for General Practice of the Radboud University Nijmegen Medical Centre were asked to participate by handing over copies of completed Compasses to the researchers. For this study, we used data of the Compass completed by GP trainers of three consecutive 3 month periods in the first year of the training.

**The Compass**

The Compass lists all seven competencies required of the general practitioner and therefore qualifies as a list of criteria that covers the entire discipline. The Compass intends to assess the seven CanMEDS competencies in an integrated, coherent and prospective way. The Compass is not an instrument that assesses trainees' competencies in one moment of time. Instead, it follows a programmatic approach\(^3\)\(^{20}\): in order to continuously gauge their trainees' competencies, trainers are to use multiple assessment instruments, on several occasions, in various working situations throughout the entire training. With the Compass GP-trainees' performance in practice is monitored and compared to the required standard throughout the training. The instrument was designed to identify trainees' performance during the training as well as the trainees' progress towards the standard of performance required upon completion of this training. The Compass enables trainers and trainees to monitor the learning process and, if necessary, remedy possible lack of performance.

During the completing of the list, the trainer:

- gives an overall assessment of each competency based on a number of indicators.
- specifies which sources of information and assessment instruments were used to gather information within each of the competencies.
- writes down considerations that have had a significant influence on the assessment per domain.
- discusses the results with the trainee during the progress review interview.

**Content analysis of the considerations about judgments**

Because of the explorative nature of this study we performed a qualitative content analysis using the constant comparative method approach\(^21\). The software program ATLAS.ti version 6.2 was used to process the data. The first step in the data analysis was data reduction by coding the transcript by two researchers (FT and TvR) independently to minimize subjectivity. Coding is the interpretive process by which data are broken down by giving conceptual labels to the data. Its purpose is to give new insights by breaking through standard ways of thinking about phenomena reflected in the data. The sequences were as follows: per competency codes were given closely related to the text fragments. We consider a code as a meaningful unit comprising sentences or part of sentences containing aspects related to each other through their content and context. After five Compasses the codes of the analysts were compared, discussed (FT, TvR, AK) until consensus was reached and a coding scheme was developed to be used for further analysis. New codes could be added. When no more new codes came forward this was an indication that
saturation had been reached. Codes referring to the same phenomenon were grouped into themes (FT, Tvr, AK). The results of the discussion were also discussed with the remaining authors (FT, Tvr, AK, Mvd, BB). In order to gain more insight into differences in elaborateness between the competency domains the completed Compasses were analyzed by counting the number of codes per competency. More codes refer to a more elaborate underpinning of the judgment.

**Results**
We analyzed 90 lists of 38 GP trainees before we reached saturation. The coding of the written comments resulted in 5 themes and two categories of competencies: Medical Expertise, Communication and Professionalism (first category) and Collaboration, Management, Social Accountability and Science & Education (second category) (table 1). The first category of competencies consisted of three themes, also the written considerations were more elaborate (table 2). Considerations of the first category appeared to be more specific, focused on performance, development and personal characteristics. In the second category we found 5 themes and the considerations were less elaborate. In the competencies of the second category, provisional judgments were given and in addition, judgments of these competencies seemed to be blurred, as other competencies were included in the considerations.

**First category**
Within the content of the considerations of this group of competencies three themes could be distinguished:
1. Judgment
2. Giving directions
3. Impressions

**Judgment**
Trainers appraised the trainees' performance and development, examples of values were 'good', adequate, improved, etc.

*Twelve minutes consultations pass off quite good* (list 0309-27-3, Communication).

*Good balance between personal and professional roles* (list 0309-22-1, Professionalism).

Often, the assessment was not based on behaviour but on trainees' personal characteristics: shy, nice person, insecure:

*Sympathetic towards patients* (list 0309-26-1 Communication).

*Insecurity fades into the background* (list 0309-14-2 Medical Expertise).

Also, at times compliment-like remarks were written down, without being clear on what behaviour the compliment was based:

*Hats off!* (List 0309-01-1 Professionalism).
Table 1. Themes, categories and codes of considerations written down by GP-trainers

<table>
<thead>
<tr>
<th>Categories</th>
<th>Themes</th>
<th>Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>First category: Competencies</td>
<td>Judgment:</td>
<td>Positive judgment specific behaviour or sub competency</td>
</tr>
<tr>
<td>Medical Expertise, Communication and Professionalism</td>
<td>of performance</td>
<td>Positive judgment competency overall</td>
</tr>
<tr>
<td></td>
<td>of development</td>
<td>Negative judgment specific behaviour or sub competencies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Negative judgment competency overall</td>
</tr>
<tr>
<td></td>
<td>of personal characteristics</td>
<td>Uncertainty</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shyness</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Attitude</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eager to learn</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Confidence</td>
</tr>
<tr>
<td>Giving directions</td>
<td></td>
<td>Needs attention</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Advise</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Instruction</td>
</tr>
<tr>
<td>Impressions</td>
<td></td>
<td>Showing initiative</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pleased easily</td>
</tr>
<tr>
<td>Second category: Competencies</td>
<td>Judgment:</td>
<td>Positive judgment specific behaviour or sub competency</td>
</tr>
<tr>
<td>Collaboration, Management, Social Accountability and Science &amp; Education.</td>
<td>of performance</td>
<td>Positive judgment competency overall</td>
</tr>
<tr>
<td></td>
<td>of development</td>
<td>Positive judgment specific behaviour or sub competency</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Positive judgment competency overall</td>
</tr>
<tr>
<td></td>
<td>Provisional judgment</td>
<td>Too little information</td>
</tr>
<tr>
<td></td>
<td></td>
<td>So far not seen</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No priority</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not visible</td>
</tr>
<tr>
<td></td>
<td>Judgment different competency</td>
<td>Involving other competencies in judgment (e.g. Communication skills, professional behaviour or Medical knowledge).</td>
</tr>
<tr>
<td>Directing</td>
<td></td>
<td>Needs attention</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Advise</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Instruction</td>
</tr>
<tr>
<td>Impressions</td>
<td></td>
<td>Prepares tasks well</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Good collaboration skills</td>
</tr>
</tbody>
</table>

**Giving directions**

Based on these judgments, trainers give advices or instructions, or formulate points of interest:

*Good structure and taking control in consultations. Upcoming months study in depth context and background of the complaints and fine-tune it with the problem of the patient (list 0309-05-1, Medical expertise)*
Chapter 4

IMPRESSIONS
Judgments seemed sometimes to be based on global perceptions and impressions of the trainees' performance in the past period:
- Shows a lot of initiative (list 0309-16 Medical Expertise).
- Sometimes to easily satisfied with the patients' story (list 0309-1 Communication).
- Shows attention for contextual circumstances (list 0309-15-2 Medical Expertise).

Table 2. Number of codes that emerged in each competency domain

<table>
<thead>
<tr>
<th>Medical Expertise</th>
<th>Communication</th>
<th>Professionalism</th>
<th>Collaboration</th>
<th>Social Accountability</th>
<th>Management</th>
<th>Science &amp; Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>101 codes</td>
<td>91 codes</td>
<td>67 codes</td>
<td>48 codes</td>
<td>45 codes</td>
<td>43 codes</td>
<td>29 codes</td>
</tr>
</tbody>
</table>

Second category
Within the content of the written considerations of this group we identified five themes.
1. Judgment
2. Provisional judgment
3. Judgment that includes other competencies
4. Giving directions
5. Impressions

JUDGMENT
In the assessment of these competencies we found no negatively phrased judgments. Only positive and/or provisional judgments were given.

PROVISIONAL JUDGMENTS
We found that trainers gave provisional judgments, although they had little or no information to go on, still judgments were given.
- As yet seen little of it. Seems to be able to develop good connections (list 0309-03-1 Collaboration).
- Cannot be judged adequately as yet. So far so good (list 0309-1-1 Science and Education).
- Seen little evidence, seems OK (list 0309-22-1 Management).
- Has not come up yet, feel good about it! (list 0309-01 Social Accountability)

JUDGMENT THAT INCLUDES OTHER COMPETENCIES
Judgments of the competencies in group 2 appeared to be blurred. Our analysis showed that trainers used competency domains interactively when judging one particular competency. When assessing Social Accountability for example, one trainer considered the communication style of the trainee to be respectful when she was giving patients advice on lifestyle issues, but did not comment on the appropriateness of the provided advice.
- Approaches patients with respect during lifestyle advice (list 0309-26-1 Social Accountability).
Discussion
This study focused on the considerations written down in the Compass about how trainers came to their judgment. Two categories of considerations emerged: Considerations regarding the competencies Medical Expertise, Communication and Professionalism (first category) and regarding Collaboration, Management, Social Accountability and Science & Education (second category). The considerations written down in the first category were more elaborate and specific, focused on performance, development and personal characteristics. In the second category provisional judgments were given and in addition, judgments of these competencies seemed to be blurred, as other competencies were included in the considerations. Management, Collaboration, Science and Education are relatively new competencies and may be harder to identify for trainers. A possible reason for this could be that they might be more familiar with the more traditionally assessed competencies of the first category. Moreover, the competencies of the second category are not confined to the consultation room and refer to additional demands following from current developments in GP practice. For the assessment of Medical expertise, Communication or Professionalism several tools are available, such as the mini-CEX, the Nijmegen Professionalism Scale and the Maas-global\textsuperscript{20,22,23}. Holmboe et al. pointed out that new assessment tools need to be developed for the 'new' competencies in order to realize the full potential of competency based medical education\textsuperscript{4}.

The weight of used instruments, or that of sources such as practice staff or fellow doctors did not become clear. This does not mean that sources and instruments did not play a part in the assessment: it only means that it was not written down. Programmatic assessment in competency based medical education faces many challenges\textsuperscript{24}. One of them is that trainers have to collect a lot information and document it, for example, in a logbook. It is of great importance that all actors in programmatic assessment should understand what they are doing, why they are doing it and why they are doing it this way. If not, they are in danger of losing sight of the true purpose of assessment and will fall back on bureaucratic procedures and meaningless artefacts. To prevent bureaucracy, support systems are needed to facilitate the entire process. Computer technology seems an obvious candidate for an important role as facilitator for documentation and storing information. The exploration of these technologies has just begun, but they show great promise to reduce workload and provide intelligent solutions to some of the problems\textsuperscript{24}.

Limitations
This study has limitations, notably the Compasses we analyzed were completed in one of the eight departments of general practice in the Netherlands. We have no data of the other departments. On the other hand, we did have a large dataset and reached saturation. Another strength is that we obtained the lists at the end of the year, and trainers completed the Compass without anticipating that their efforts were used for research. This may have reduced bias.
Implications of our study
Feedback plays an important role in competency based medical education. Our results indicate that providing feedback on the competencies Management, Collaboration Science and Education and Social Accountability might be difficult for trainers. Feedback is often based on observation, and these competencies may not be clearly visible in daily practice. Therefore, procedures to assess the competencies Management, Collaboration Science and Education and Social Accountability are needed. Not only assessment tools but also a training should be developed so that professionals can learn to use the newly developed tools.

Further study
Trainers did not write down, which sources and instruments plaid a significant part in the assessment of the seven competencies. Further research may clarify how trainers use these sources and instruments and how they aggregate the information they have gathered in a predetermined training period. Also of interest is how trainers document their information. For instance, do they make use of a logbook or do they place the information in some sort of file?

Conclusion
Two categories of competencies could be distinguished. The written content of the perhaps more familiar competencies was more extensive and elaborate than the content of the 'new' ones. We have speculated on a number of reasons why. Further study should result in finding appropriate ways to assess the competencies Collaboration, Management, Social Accountability and Science & Education.
References

Chapter 5

Development of a multi-method selection procedure for postgraduate training based on the CanMEDS in the context of GP training

F. Tromp
M.I. Vermeulen
H.G.A. Mokkink
M.J.F.J. Vernooij-Dassen
B.J.A.M. Bottema
A.W.M. Kramer

Submitted.
Abstract

Introduction: Most specialties select trainees on an intuitive and often poorly defined basis. Our aim was to present the development of a fair and standardized selection procedure for postgraduate training based on empirical evidence.

Methods: First, the content of the procedure was determined with a modified Delphi procedure. Then, we determined which instruments should be used in the procedure.

Results: Consensus on the assessment of the following CanMEDS competencies was reached: Medical Expertise, Communication, Collaboration, Management, and Professionalism. Four instruments were included: a knowledge test; a situational judgment test; a series of three work-related simulations; and a competency-based interview.

Conclusion: A competency-based multi-method selection procedure for postgraduate training based on empirical evidence and the CanMEDS framework is available for our training. This procedure will allow for comprehensive standardization that is fair to the candidates. The results can be used during training as a baseline assessment by trainers and candidates. They can use the scores of the selection instruments to identify future development. Further research is needed to establish the reliability and predictive validity of the procedure. Other medical specialties that utilize the CanMEDS or comparable competency frameworks as a basis for their curriculum could employ this stepwise development model.
Introduction

Selection procedures play a crucial role in obtaining access to medical postgraduate training. These procedures should be credible, fair, and publicly defensible. For many occupational groups, a large body of international research exists investigating best practice selection\(^1\). In medicine, there is a significant volume of research exploring medical school admission procedures and the link to subsequent performance during medical school. There is relatively little research on developing selection methodology for entry as a trainee to postgraduate training\(^2,3\). As a result, most specialties continue to select trainees on a subjective and often poorly defined basis\(^4\). To improve the selection process, Prideaux et al. postulate that selection should be conceptualized as "assessment for selection"\(^5\). In doing so, the well-developed quality assurance mechanisms associated with high-stakes assessment can be applied in the selection process.

The first quality assurance mechanism is "proceeding from a clear blueprint of the content for selection"\(^3\). The importance of a thorough blueprint has been underlined previously\(^5,6\). At present, competency frameworks, such as those developed by the Accreditation Council for Graduate Medical Education and the American Board of Medical Specialties (ACGME/ABMS)\(^7\) and the Canadian Medical Education Directives for Specialists (CanMEDS) 2000\(^8\), guide the construction of curriculums in many countries. These competency frameworks constitute the essential abilities that physicians need for optimal functioning\(^8\). The implementation of these competency frameworks may have consequences for selection, as the competencies can be used for the content of an assessment procedure. They provide all of the qualities for which a specialist should strive as a medical expert, communicator, collaborator, scholar, manager, health advocate, and professional. Not all competencies are appropriate for selection. Some subcompetencies, such as showing empathy, are difficult to teach, and should have been developed already during medical school. Other competencies, such as the execution of specific surgical procedures, can be developed more easily during postgraduate training. This distinction is crucial for determining the selection criteria for each specialty.

The second issue emphasized by Prideaux et al. is that selection should be aligned with the curriculum and assessment. The majority of selection procedures for postgraduate entry are based on cognitive variables, and these variables alone do not adequately predict the performance of competencies such as Professionalism, Communication, or Management\(^3,5,6,9\). Assessing these competencies during selection may actually be more predictive for success in a residency than traditional cognitive selection factors. By taking the competency framework as a starting point and assessing the same competencies as in training, the selection procedure is more aligned with curriculum and assessment.

In conceptualizing selection as "assessment for selection", we can benefit from the findings of experts in the field of assessment and apply these findings to the assessment for selection\(^5\). Van
der Vleuten and Schuwirth state that one instrument is not sufficient in high-stakes assessment\textsuperscript{10}. This insight inspired them to advocate programmes of assessment. In a programme of assessment multiple sources of information from various methods are used to construct an overall judgment by triangulating information across these sources. Likewise, the selection procedure should consist of several assessment tools and should be considered as the first assessment in a programme of assessments. The results of the selection procedure can be seen as a baseline assessment. Such a selection procedure enables future trainees to receive feedback at the very beginning of the training. Assessments should generate feedback, because feedback promotes learning: it advises trainees regarding observed learning needs; and it motivates trainees to engage in appropriate learning activities\textsuperscript{11}.

There have been important developments in the domain of selection for postgraduate training. In the UK, Patterson et al.\textsuperscript{12} developed a new selection procedure based on competencies to recruit general practice (GP) trainees. This procedure showed predictive validity for future performance during the training of candidates. In the Netherlands, Vermeulen et al. studied the current selection procedure of GP training\textsuperscript{13}. This selection procedure is endorsed nationally but conducted locally. It was found that despite the legislation, different standards were used in different institutes and the department itself was a predictor of being admitted. Viewing the results of their study, the authors expressed their doubts about the fairness of the selection procedure and suggested that the current method be reconsidered.

Our aim was to develop a fair, standardized selection procedure for GP training based on empirical evidence and on the leading theoretical studies. Although this procedure was developed in the specific context of GP training, other specialties can benefit from our experiences by describing the process of the development.

**Methods**

**Context of the study**

We conducted our study in Dutch Postgraduate Training for general practice. This training has a nationally endorsed curriculum. The GP departments of the eight university medical centers are responsible for the organization of the three-year postgraduate training. The curriculum is based on the CanMEDS competencies\textsuperscript{8}, which are adapted to the specific needs of the specialty. These competencies are assessed during training with the Competency Assessment List (‘Compass’), an instrument that lists the seven competencies\textsuperscript{14}. The Compass aggregates the assessments of performance in practice at several points during training. At the time we developed the selection procedure, the number of candidates was decreasing, while the number of vacancies/places increased. In order to maintain quality of the training, we decided to ‘select out’, meaning that we aimed to identify unsuitable candidates.
Dropouts and poor performers, even if their number might be small, cost the departments substantial effort and money. Their places cannot be filled, which is a waste of resources. We aimed to develop a selection procedure for postgraduate GP training that

- is able to identify unsuitable candidates;
- is preceded by a content analysis and based on relevant and actual competencies for the specialty;
- exhibits congruity between selection, curriculum, and assessment;
- uses multiple assessment instruments on different levels of Miller with satisfactory predictive validity and reliability; and
- is feasible: the whole procedure has to be executed in not more than one day.

**Design of the study**

Our study consisted of two steps:

**Step 1: Establishing the content of the selection procedure**

To determine which of the CanMEDS competencies should be targeted for the selection procedure for GP training, we invited a panel of 16 experts, all involved in selecting candidates for GP training. In a two-round process, we asked the panelists to judge "which of the CanMEDS-competencies should already be present before entering GP training to finish the training successfully". This inquiry indicates that we aim to identify unsuitable candidates with the new procedure. In cases where one or more of the CanMEDS competencies are missing, candidates will not be admitted to the training, i.e., they will be 'selected out'.

We used the Compass format to determine the content because it lists all seven competencies and their 19 subcompetencies (box 1). The panel members individually rated the subcompetencies on a nine-point scale, ranging from should not be present before entering GP training (=1) to should certainly be present before entering GP training (=9). In the first round, the ratings were made individually at home, with no interaction among the panelists. In the second round, the panel members met under the leadership of a moderator. During the meeting, the panelists discussed their previous ratings, focusing on areas of disagreement. Disagreement among the raters was defined if at least one third of the panel members rated the subcompetency in the range of one to three while at least one third of the other panel members rated the subcompetency treatment in the range seven to nine. After discussing each of the seven competencies, they rerated each subcompetency individually. The two-round process was focused on detecting consensus among the panel members. No attempt was made to force the panel to consensus. The subcompetencies were accepted when they received a mean score of seven or higher.
To corroborate the results of the expert panel and obtain additional information, we organized a focus group meeting with the heads of the training departments. These individuals have a good impression of the problems of non-functioning trainees and dropouts, as they are responsible for the pass-or-fail decisions. The heads of the departments were asked which of the competencies most frequently cause the most significant problems. The panels were not informed about the results of the other panel, and we combined the results of the two panels.

**Step 2: Determination of assessment tools**

The second step was to determine which instruments should be used to assess the competencies that we found during the first step. In a high-stakes situation, no single assessment instrument can provide the necessary information for judgment\(^{10,15}\). We chose to include various instruments that provide different levels of information. Corresponding to the classification that George Miller proposed for the methods of assessment in medical education\(^{15}\), we aimed to assess the factual knowledge and performance of the candidates. We assessed whether the candidates were able to act appropriately in a practical situation by exhibiting functional behavior. This approach complements our endeavor to achieve congruity between the selection, curriculum, and assessment because during training trainees are assessed in a similar manner.

The included instruments should have a good predictive validity and reliability, as confirmed by the literature. We conducted a search using the PubMed database. The following search terms were applied: "internship and residency", "education, graduate", "vocational education", "school admission criteria". We used the database Psychinfo applying the keywords "personnel selection" with the limitation "meta-analysis".

**Results**

**Step 1: Establishing the content of the selection procedure**

Of the 16 individuals that we approached, 11 were willing to participate. The panel reached consensus on eight of the 19 subcompetencies in the first, written rating round. In the second round, two subcompetencies of Medical Expertise, one of Communication, three of Collaboration, two subcompetencies of Management, one of Social Accountability, and two of Professionalism were discussed. After hearing positive and negative arguments, the panel re-rated the subcompetencies at the end of the meeting, resulting in consensus on five competencies and nine of their respective subcompetencies. The heads of the departments reported that the competencies Medical Expertise, Communication, Collaboration, and Professionalism caused the greatest difficulty during training. They felt that a significant lack of medical knowledge was an important cause of problems or dropping out during training.

The results of the panel of experts and the heads of the departments overlapped to a large extent, and the only competency that was not reported by the department heads was
Management, which was contrary to the experts. In box 1, the targeted competencies are printed in italics.

Box 1. **Overview of the seven competencies**

<table>
<thead>
<tr>
<th>Competencies</th>
<th>Subcompetencies</th>
</tr>
</thead>
</table>
| **Medical Expertise** | *Interprets symptoms in context*  
*Applies the diagnostic, therapeutic, and preventive arsenal of the profession in an appropriate and evidence-based manner*  
Provides primary care in a systematic manner |
| **Communication** | *Develops effective treatment relationships with patients*  
*Applies communication techniques and resources appropriately*  
Ensures that the patient is actively involved in the decision making |
| **Collaboration** | *Contributes to effective intra- and interdisciplinary collaboration*  
*Applies collaboration skills appropriately*  
Makes appropriate referrals on the basis of a current insight into the expertise of other care providers |
| **Management**   | *Provides integral and appropriate general practice care that is continuous and accessible*  
*Applies organizational and management principles appropriately*  
Uses information technology for optimal patient care |
| **Social Accountability** | Promotes the health of individual patients and groups of patients  
Acts in accordance with the legislation that applies to the general practitioner |
| **Science and Education** | Underpins care in an academically sound manner  
Promotes the expertise of students, trainees, colleagues, and other care providers |
| **Professionalism** | *Maintains a balance between personal and professional roles*  
*Works systematically and purposefully to improve his or her professional performance*  
*Addresses differences in standards and values consciously within the context of professional ethics* |

Subcompetencies that are targeted for selection are printed in italics.

**Step 2: Determination of assessment tools**
For practical reasons, four instruments were included because we felt that additional instruments made the procedure overly lengthy. The format for each instrument, which describes the competencies that are assessed and evidence from the literature, will be discussed.

1. **The knowledge test for general practice**
Because the heads of the departments felt that a significant lack of medical knowledge was an important cause of problems or even dropping out during training, we decided, also for pragmatic and cost efficiency reasons, to assess medical knowledge with a validated knowledge test for GP\textsuperscript{16,17} that is used during training to assess the progress of knowledge. The knowledge test is based on a blueprint covering all seven competencies\textsuperscript{17} and consists of 120 questions with "correct"/"incorrect"/"do not know" answers. To discourage guessing, the overall score is calculated as the sum of the correct minus incorrect answers and is expressed as a percentage of the maximum score.
EVIDENCE FROM THE LITERATURE
Schmidt and Hunter have reviewed the literature on the predictive validity of the various selection instruments and found high predictive validity for knowledge tests. In the medical domain, it was shown that medical knowledge is the basis of performance in practice. Knowledge tests have high predictive validity on future performance. In the medical domain, it was shown that medical knowledge is the basis of performance in practice. In general reliability of the LHK varies between 0.60 and 0.76.

2. Situational judgment test
To assess the ability to use knowledge in a particular context, a situational judgment test (SJT) was included. In a SJT, the candidates are presented with written depictions of professional dilemmas that they may encounter in practice and are asked to identify an appropriate response from a list of alternatives. This test assesses the competencies: Professionalism, Management, Collaboration, and Communication.

With the Critical Incidence Technique (CIT) experienced GP's formulated 20 professional dilemmas for the SJT, each with four alternatives. The alternatives can be rated as "very appropriate", "appropriate", "neutral", "inappropriate", and "extremely inappropriate". An example is provided in box 2.

EVIDENCE FROM THE LITERATURE
The studies of Lievens and Patterson have demonstrated the good predictive validity of the SJT. In a meta-analysis it was found that the reliability of SJTs ranged from 0.43 to 0.94. In selection for postgraduate training in general practice in the UK, internal consistency of the SJT ranged from 0.80 to 0.83.

Box 2. Example of a professional dilemma in the SJT

A man visits his GP. He is feeling very tired lately. He has been working very hard, but he is concerned something serious is the cause of his complaints. His GP has examined him and concludes that his symptoms are normal in these circumstances.

Patient: I don't trust this, can't you arrange a more extensive examination?

GP: 1. I could do that, but I am telling you that your complaints are normal, so why look any further? It won't gain you anything if we do that.
2. OK, if you are still worried and I can do nothing to take away your concern, maybe it is for the best to put in a request.
3. You're still worried? OK, let us find out the reason why. Maybe I can reassure you.
4. You're still worried! I don't understand why. I think you are somewhat exaggerating.

3. Simulation exercise
To assess the candidate's ability to act appropriately in a practical situation, a simulated situation exercise was included. The simulations provide the candidates with a good impression of daily
practice. The competencies Medical Expertise, Communication, Management, Collaboration and Professionalism are assessed with this exercise.

The scripts were developed also using the CIT, and an example is provided in box 3. This exercise applies the same principles as the Multiple Mini Interview (MMI) and the Objective Structured Clinical Examination (OSCE). Both MMI and OSCE provide a series of short testing stations and have shown to have superior reliability to a single long case. The simulation exercise employs a series of short scenarios with observers in each station or scenario. As with the OSCE and the MMI, the simulation exercise diminishes the problem of context specificity, in which the measurement of an attribute in one context does not necessarily transfer to another. To reduce the context specificity, we developed three short simulations. The observers participated in one day of training in behavioural observation and rating.

EVIDENCE FROM THE LITERATURE
Work-related simulation exercises exhibited high predictive validity. The MMI format shows satisfactory levels of reliability, content- and predictive validity.

Box 3. Example of a script of the simulation exercise

<table>
<thead>
<tr>
<th>Instruction:</th>
</tr>
</thead>
<tbody>
<tr>
<td>You are a GP working in a group practice. The surgery hours are very busy today. You are already overrun by 30 min. The next consultation is with Mrs. R. She has been a patient in your practice for only one year, so you don’t know her very well. Mrs. R indicated that her reason for the encounter is headache. This instance is not the first time she consults you for this reason. From her medical file, you know she saw a neurologist 18 months ago. She was diagnosed as having pain that originated from muscle tension. She has had physiotherapy with varying success. Mrs. R is married and has two children. Remember: you have only 10 minutes for the consultation. The waiting room is full, and you are already 30 min late. Your assistant has urged you to hurry and make up for lost time because one of your colleagues had to leave suddenly for personal reasons. You have to take over some of his patients as well. Your assistant also warned you that Mrs. R can be very long-winded. Physical examination is not necessary. You may assume that a physical examination will not yield further information.</td>
</tr>
</tbody>
</table>

4. Structured interview
The fourth assessment instrument that we chose was the Patterned Behavior Description Interview (PBDI), which is widely used in selection. The PBDI is based on the premise that past behavior predicts future behaviour. The interview focuses on the evaluation of reactions in actual situations from each candidate's past, relevant to the targeted competencies. The assessed competencies are Communication, Management, Collaboration, and Professionalism. The interviewers were trained in the technique of this specific format of interviewing and how to rate the candidates' answers.
Chapter 5

Evidence from the Literature

Schmidt and Hunter have demonstrated that good validity coefficients are found for the more structured and systematic techniques, such as structured interviews\(^1\). In the medical literature, we found that the use of behavior-specific questions during the interview improved the predictive validity\(^{19,33,34}\).

In summation, our procedure consisted of

- The Knowledge Test for General Practice
- An SJT
- A simulation exercise
- A structured interview

Table 1 summarizes which competencies are assessed by each of the four instruments.

<table>
<thead>
<tr>
<th>Medical Expertise</th>
<th>Knowledge Test</th>
<th>SJT</th>
<th>Interview</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interprets symptoms in context</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applies the diagnostic, therapeutic, and preventive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>arsenal of the profession in an appropriate and</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>evidence-based manner</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communication</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develops effective treatment relationships with</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>patients</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applies communication techniques and resources</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>appropriately</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collaboration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applies collaboration skills appropriately</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applies organizational and management principles</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>appropriately</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professionalism</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintains a balance between personal and professional</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>roles</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Works systematically and purposefully to improve one's</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>professional performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Addresses differences in standards and values</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>consciously within the context of professional ethics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discussion

A competency-based selection procedure for postgraduate training based on empirical evidence and theoretical insight is now available for postgraduate training in GP. We developed this procedure guided by empirical data\(^2,6,9,12,35\) and the recommendations of the leading experts in the field\(^3\).
This procedure allows for comprehensive standardization that should be more fair to the candidates than the existing methods. To ensure good standardization in the future, all of the GP departments should work together to produce procedural and best practice manuals and ensure that all assessors and simulators have been properly trained for the best practice selection procedures, as is the case in the UK. The involvement of all departments is crucial and was the key factor for success in the UK in developing a standardized national selection procedure.

To our knowledge, the developed selection procedure is the first to be based on the CanMEDS. One of the characteristics of the procedure is a job analysis to classify the core and specific competencies, as recommended by Prideaux. With the aid of experts in the field, we decided which competencies should be targeted for selection. These generic formulated competencies were translated to behavioral indicators using the CIT. By assessing the candidates according to the content of the curriculum, the candidates are confronted with what will be expected of them during training. With this transparency, the candidates are likely to develop a realistic perception of the job role. This assessment may potentially reduce the number of false positives, thereby reducing attrition rates and problems occurring during training because candidates would have a more realistic insight of what the job entailed before enlisting.

To determine the predictive validity of various instruments, we found that the results were not univocal. There seems to be a considerable debate about the predictive validity of each single instrument. In our procedure however, we do not use one single instrument. No single instrument can assess all competencies. Moreover, one single assessment has limitations, such as case-specificity, or low reliability. The four instruments we chose provide information of the candidates on different levels. Our procedure will incorporate several competency elements and multiple sources of information to evaluate those competencies on multiple occasions using credible standards. The information obtained will have to be aggregated into a final decision.

Trainers and trainees can use the scores of the selection instruments to identify future development needs for candidates. The procedure can be considered a baseline assessment. It enables future trainees to receive feedback before the training begins. With the aid of this feedback, the future trainees are able to remedy potential shortcomings in the earliest stage of the training. The ability to provide and receive feedback at such an early stage of training is unique. The rejected candidates will receive feedback enabling them to work on their deficiencies, thus giving them a fair chance if they choose to apply again.

Although our study was based on developments in the UK, our procedure has one important difference. Although interviews are the most common selection method, interviews are not
included in the selection procedure in the UK. We deliberately included the PBDI because it is widely used in selection and has good predictive validity¹.

Having established a theory-driven and evidence-based selection procedure, the next step will be to evaluate its effects and costs. The reliability and predictive validity of the proposed procedure will be assessed by conducting a longitudinal study, and the candidate's perspective and perceptions of fairness must be considered³⁷. Further work exploring how best to utilize the output from the selection process to inform personal development planning is important.

Limitations of the study
High numbers of candidates for places in medical training could prove to be a substantial investment of time and money and so jeopardize feasibility. One solution to unburden the procedure could be to use the Knowledge Test and the SJT as preselection tools, as in the UK²,³¹. We did not include former academic performance as a selection variable. Although there is some criticism among researchers that the educational curricula and quality of teaching may differ among institutions³⁸, academic performance can be considered as a good predictor of future performance⁵,³⁹,⁴⁰; this consideration will be investigated in the future.

Conclusion
We developed a selection procedure based on the CanMEDS using selected valid instruments assessing multiple competencies. The results of the selection procedure can be used at the start of the training as a baseline assessment. The next step will be to study the instrument's reliability, predictive validity fairness, costs, and future use in training, as well as the candidates' reactions.

This study described a stepwise process for the development of a competency-based selection procedure. Other medical specialties that utilize the CanMEDS or a comparable competency framework as a basis for their curriculum could employ this development model.
References


Chapter 5

Chapter 6

Summary and general discussion
Chapter 6

Introduction
In this thesis we focus on the development and validation of competency-based assessment procedures that can evaluate performance at the workplace, covering a broad scope of competencies and giving insight in learner's development. The development of these procedures was guided by theoretical models and frameworks of assessment, specifically Miller's classification of clinical performance\(^1\), frameworks of workplace-based assessment as proposed by Pangaro and Ten Cate\(^2\) and the conceptual model of programmatic assessment of Van der Vleuten\(^3\).

Our focus was on the content, on the reliability, on the feasibility and on the process. We conducted this study in the 'real world' of GP-training and had to adapt our study to the ongoing educational developments that were implemented.

In this chapter we will discuss the main findings, how to interpret these findings in relation to the literature, the methodological considerations, what this study adds, future research, the practical considerations and the conclusion.

Main findings
The most important finding is that the Compass is the first competency-based instrument that successfully demonstrates progress of performance across all competencies of GP-trainees. The results of our study indicate that by using a fixed-reference point for trainers' judgments, which is the standard expected by the trainer upon completion of the three-year training, it is possible to visually review progress over time.

Next, we present the main results of each of the four studies.

Nijmegen Professionalism Scale
The Nijmegen Professionalism Scale consists of four themes of professionalism: professional behaviour towards patients, other professionals, society and oneself. To examine the construct validity of the Nijmegen Professionalism Scale, we conducted a factor analysis and failed to replicate the instrument's original structure. Our factor analysis revealed a much simpler structure with two dimensions for each theme. In "professional behaviour towards the patient" we labeled the factors as respecting patients' interest and professional distance. In the theme "professional behaviour towards other professionals" the factors were labeled as collaboration skills and management skills, in the third theme "professionalism towards society" we named the factors responsibility and quality management. Finally in the fourth theme of the instrument, "professionalism towards oneself" we labeled the factors reflection and learning and dealing with emotions. In addition, we compared the factor structure of trainees' self-assessments with the trainers' evaluations. No considerable differences in the four domains were found. This indicates that GP trainers and trainees attached similar meanings to the construct of professional
behaviour, creating a solid foundation for the effective teaching and assessment of this essential part of medical performance. In our view, the revised structure improved the conceptual clarity of the instrument and rendered it more straightforward for GP trainers to use.

Cronbach's alpha coefficients ranged from 0.79 (dealing with emotions) to 0.95 (reflection and learning), indicating good to excellent internal consistency within each dimension\(^4\).

Further information supporting the validity of the Nijmegen Professionalism Scale could come from data suggesting that it can accurately identify trainees with performance deficits and that the instrument can also measure their professional growth. Because we did not detect ceiling effects, growth can, in principle, be assessed, but we did not specifically test this.

The ability of the NPS to measure professional growth of GP-trainees was not tested because at that time the Compass was implemented in GP-training. The Compass included behavioural elements of the Nijmegen Professionalism Scale and also applied its format of longitudinally assessment.

**The Compass**

The Compass follows a programmatic approach by aggregating separate assessments of performance in practice\(^{3,5}\). To continuously gauge their trainees' competencies, trainers are to use multiple assessment instruments on multiple occasions and in various working situations throughout the entire length of training. Every three months, the aggregated results of these assessments are summarised in the Compass and are then discussed between the trainer and trainee.

The first version of the Compass was developed by consensus among experienced GP's and medical educators in general practice training according to the CanMEDS guidelines. Our validation study, using a RAND Delphi consensus procedure with GP-trainees and GP-trainers, revealed that only minor changes to the content of the Compass were needed.

All scales showed excellent internal consistency ranging by Cronbach’s alpha ranging from .89 to .94.

The Compass demonstrated the progress of performance in daily practice for all competencies of the GP-trainees. Progress was visualised as an increase in rating over time. To gain insight into how trainers form their judgments of trainees' performance and progress, we conducted a qualitative study. This study focused on the considerations regarding how trainers formed their judgments written down in the Compass. Two categories of considerations emerged: considerations regarding the competencies Medical Expertise, Communication and
Professionalism (first category) and considerations regarding Collaboration, Management, Social Accountability and Science & Education (second category). The notes recorded in the first category were more elaborate and specific, focusing on performance, development and personal characteristics. In the second category, provisional judgments were given and, in addition, judgments of these competencies seemed to be blurred, as other competencies were included in the considerations.

**Selection procedure**

We developed a standardized selection procedure based on empirical evidence and on theoretical studies\(^3,6-10\). Moreover, we conceptualized selection as "assessment for selection"\(^6\). We started with a job analysis, using the CanMEDS competencies. The following competencies were ultimately included in the procedure: Medical Expertise, Communication, Collaboration, Management, and Professionalism. Next, we decided to use several instruments that could provide information of the candidates on different competency elements and by multiple sources. Based on the literature and on feasibility characteristics we included four assessment instruments: a knowledge test, a situational judgment test, a series of three work-related simulations and a competency-based interview. Because the content of the assessments corresponds with the content of the curriculum, trainees and trainers can use the results to identify development needs at the very beginning of training.

**How to interpret these findings in relation to the literature**

To examine the construct validity of the Nijmegen Professionalism Scale, we conducted a factor analysis. Initially, the structure was based on the consensus of experts in the field of professionalism. Our factor analysis revealed a much simpler structure with two dimensions for each theme. In our view, the revised structure improved the conceptual clarity of the instrument and rendered it more straightforward for GP trainers to use. In addition, we compared the factor structure of trainees’ self-assessments with the trainers’ evaluations. No considerable differences in the four domains were found. This indicates that GP trainers and trainees attached similar meanings to the construct of professional behaviour, creating a solid foundation for the effective teaching and assessment of this essential part of medical performance. We adapted the Nijmegen Professionalism Scale according to the new structure because this is in line with how trainers actually use the list, which bolsters its validity and reliability. The results underline the importance of involving actual users in the development of assessment instruments. A similar conclusion was reported by Crossley and Jolly, who also emphasized that in workplace based assessment we should make the best possible use of the expertise of observers\(^11\). Forcing them into the straitjacket of an imposed content and structure will not benefit trainees’ assessment of professional behaviour.
In our study into the Compass, we deliberately choose completion of training as a fixed reference point. The results indicate that end of training provides clarity to the trainers about how observations of performance should be analyzed and compared to a standard of what is to be achieved by the trainee. Frameworks for workplace-based assessments are needed to give clarity to trainers and trainees about what is to be evaluated in practice, what kind of observations or assessments are useful and how performance is analyzed and compared to a standard of what is to be achieved\(^2\). Each formal performance assessment must be accompanied by an explicit criterion for determining whether or not a trainee has or has not attained the required level\(^12,13\). Methods using global ratings for observed performance often have problems such as gauging the reference point against which evaluators judgments measure standards. Prescott-Clements and colleagues argued that instruments that do not use the end of the training as a reference show little difference in scores during training\(^14\). They developed an instrument that longitudinally evaluated performance (LEP) in dental vocational training. It was found that despite initial concerns, a great deal of support was given by trainers for the use of a fixed reference point for judgment. More than seventy percent of trainers agreed that this helped drive the training. Over ninety percent thought that it helped in identifying progress over time. There are few studies that demonstrate progress of performance in daily practice during medical training. Kramer et al., for instance, did not succeed in demonstrating the growth of GP trainees’ communication skills during postgraduate training in general practice, although trainees more confidently utilized communication skills at the end of the training than at the beginning\(^15\).

In a qualitative study we found that considerations regarding how trainers came to their judgments were more elaborate for the traditional competencies Medical Expertise, Communication and Professionalism, while the remaining competencies appeared to be more difficult to address. Trainers might be more familiar with these traditional competencies. Management, Collaboration, Science and Education are relatively new competencies and may thus be harder to identify. Moreover, these new competencies are not confined to the consultation room and refer to additional demands following from current developments in GP practice. For the assessment of Medical Expertise, Communication or Professionalism, several tools are available, such as the mini-CEX, the Nijmegen Professionalism Scale and the Maas-global\(^16,17\). Holmboe et al. noted that new assessment tools for the new competencies are needed to realize the full potential of competency-based medical education\(^18\). The competencies Collaboration, Management, Organization and Health Advocacy will be increasingly relevant in primary care in view of ongoing societal changes and the shift towards primary care\(^19\). From the literature, we know that GP trainees reported that the competencies Organization and Health Advocacy were not given sufficient attention during postgraduate training\(^20\). Our results emphasize the need of more attention to the new competencies.

In postgraduate training of Dutch GPs the problems with the assessment of the new competencies has led to the development of practice-based assignments. In these assignments
the complexity of daily practice is reflected, where competencies are intertwined. With these assignments several competencies can be assessed at the same time. The approach of practice-based assignments is comparable with that of the entrustable professional activities (EPA’s) developed by Ten Cate and colleagues. These authors state that the separation in the assessment of the seven competencies is artificial and should be taught and assessed in an integrated manner. The application of practice-based assignments seems promising but its usefulness for GP-training has not been evaluated yet.

In competency-based education, with its focus on personalized, self-directed learning, the need for baseline assessment is apparent.

Starting trainees form a heterogeneous group. All individuals have different levels of competence due to differing earlier working experiences. There is some concern that a portion of the medical school graduates is not adequately equipped to meet the standards needed for entering postgraduate medical education. This discrepancy is not unique for the Netherlands. In Denmark, for instance, educators have noted that on average only 75% of postgraduate trainees have mastered the clinical skills expected for medical school graduates. In order to assess the competencies at baseline, one can consider the selection procedure to be the first assessment of the trainee. To improve the quality of selection, we conceptualized selection as "an assessment for selection". In doing so, we made use of the "best practice" principles of assessment of professional competence. One of these principles is that one instrument alone is not sufficient in high-stakes assessment. Thus, in our selection procedure, which is a high-stakes assessment, decisions are based on the outcomes of various assessment instruments.

Another best practice principle of assessment is to proceed from a clear blueprint that is generated from the content. To our knowledge, the selection procedure developed in our study is the first to be based on the CanMEDS. This competency framework constitutes the essential abilities that physicians need for optimal functioning. However, some competencies, such as showing empathy or the ability to reflect over one’s behaviour are difficult, if not impossible, to teach and should have been developed already during medical school. Other competencies, such as the execution of specific surgical procedures, can be developed more easily during postgraduate training. This distinction is crucial for determining the selection criteria. Our study indicates that with these deliberations in mind it is possible to chose, with the aid of experts in the field of GP training and selection, these competencies from the CanMEDS framework that are relevant for selection.

Moreover, by taking the competency framework of the curriculum as the basis of selection and by assessing the same competencies as in training, the selection procedure is in line with the curriculum and assessment programme. Therefore, the selection process can be considered as a
baseline assessment of training and as the starting point of the programme of assessment. One of the aims of a competency-based assessment is the guidance and monitoring of the trainee’s professional development in order to support self-directed learning. Feedback is a core component of assessment, it is central to learning and at ‘the heart of medical education’\textsuperscript{25}. In this respect, assessment and learning are closely linked to each other. Feedback also plays an important role in our selection procedure. Trainers and trainees can use the scores on the selection instruments to identify future development needs for candidates. Additionally, with the aid of this feedback, future trainees will be able to formulate learning goals to remedy potential shortcomings in the earliest stages of training. The ability to provide and receive feedback at such an early point in training is unique, and as such, this selection procedure benefits institutions as well as future trainees.

**Methodological considerations**

This thesis describes several studies, all of which contribute to the central theme of how the new approach of competence-based assessment functions in postgraduate training. The validity of newly developed procedures for the assessment of trainees’ competencies was studied to discover if they met the desired standards. We further investigated the psychometric properties of instruments that are used during training. This approach has the merit of providing evidence from the real world, thereby enhancing the relevance of the findings\textsuperscript{26}. The main strength is its relevance to the practice of medical education. Postgraduate training primarily occurs in the workplace where deeper learning depends on formative assessment, feedback and reflection. The use of qualitative methodology enabled us to acquire information that we could not obtain with quantitative methods. In particular, the results of our qualitative study gave us more insight into how trainers completed the lists. However, the results also generate new questions that are discussed in the next section.

Some limitations of the research should also be considered. Because the studies were conducted in one of the eight departments of general practice training in the Netherlands, the generalizability is not entirely certain. Moreover, although the studies address generic themes in medical education, the specifics are associated with a particular setting of GP training. This setting essentially differs from medical specialty training in that trainees are supervised by one trainer during a one-year training period.

In addition, a possible drawback of longitudinal progress assessment using rating scales is that raters may anticipate that trainees’ scores will increase with time, which could positively bias ratings awarded later in training\textsuperscript{14}. For this reason, it was interesting to see that a standstill or even a decrease in the mean ratings were observed in all CanMEDS competencies in a number of trainees.
Finally, in our studies of the Compass we only collected data from the first year of training; consequently, we do not know if the Compass is sensitive enough to measure progress in the second and third years.

What this study adds
We studied procedures that longitudinally focus on the progress of performance in practice. A major finding is that the validated Compass, using the standard upon completion of the three-year training programme as a fixed reference point for judgment, demonstrates progress of GP-trainees performance in all the competencies.

Research on assessment in medical education has accentuated individual instruments that measure performance during a single patient encounter\textsuperscript{27}. The Compass aggregates and integrates assessments made in a period of three months. This continuous way of assessment proves to be sensitive to change. By incorporating selection in the assessment programme, assessment can be employed from the very beginning until the end of the training.

Future research
In chapter 5, we described a stepwise process for the development of a competency-based selection procedure. The next step should be to study the procedures' psychometrics. After all, in a selection procedure, which is a high-stakes assessment, decisions should be based on the outcomes of reliable and valid assessments.

In chapter 4, we found that trainers did not write down which sources and instruments played the most significant roles in the assessment of the seven competencies. Further research may clarify how trainers use these sources and instruments and how they aggregate the information they have gathered in a predetermined training period. Also of interest is how trainers document their information. For instance, do they make use of a logbook or do they place the information in some sort of file?

Feedback plays an important part in competency-based medical education. Our results indicate that providing feedback on the competencies Management, Collaboration Science and Education and Social Accountability may be difficult for trainers. Feedback is often based on observation, and these competencies may not be clearly visible in daily practice. Therefore, procedures for the reliable and valid assessment of Management, Collaboration Science and Education and Social Accountability are needed. However, assessment tools are not sufficient. In workplace-based assessment the users play an important role. Therefore, training should be developed so that professionals can learn how to use these new tools. Moreover, we need more information about what the effect is of feedback coming from these assessments. Do trainees formulate more
learning goals? How specific are these goals? What do they do to achieve them and what is the effect on performance?

We did not study the use of the Compass in the second and third years of training. It is possible that progress may diminish during those years. Kramer showed that medical knowledge, after significant growth in the first year, stagnated in the second year of training and grew again in the third year, albeit not as much as during the beginning of training28. Does progress in performance show the same pattern and is the Compass sensitive enough to detect progress as training continues?

**Practical recommendations**

We firstly recommend the use of a fixed reference point in the assessment. The desired standard upon completion of the three-year training gives a clear picture to both trainee and trainer of what is to be expected. Typically, trainers often report significant uncertainty about the standard expected at a given stage of training14.

Competency-based assessment in postgraduate medical training is a work in progress, so many changes must be implemented. One of the conditions of successful implementation is that the leading individuals of an organization consent and facilitate changes that are based on scientific evidence29. The results of our study of the Compass were published in 2012. However, in 2013 a nationally endorsed guideline for assessment was introduced that stated that the reference point for trainers' judgments should be the stage of training instead of the standard expected on completion of the training. This guideline was agreed upon despite the results we presented. We feel that this is a missed opportunity for GP-training.

We recommend to involve trainers in the development of workplace-based instruments. After all, they are familiar with the daily practice, and early involvement may improve the validity, reliability and implementation of the instruments, as they are better tailored for practice.

We described a stepwise process for the development of a competency-based selection procedure. Medical specialties that utilize the CanMEDS or a comparable competency framework as a basis for their curriculum can employ this process for the development of a selection procedure for their specialty. To align selection, curriculum and assessment, the same competencies are assessed during selection and training. Therefore, the results of the selection procedure can be used at the start of training as a baseline assessment.
Conclusion
In this thesis, we studied assessment procedures that evaluate performance at the workplace, covering a broad scope of competencies and giving insight in learner's development. This study was conducted in 'the real world' and we faced many changes along the way. However, we built on our experiences by using the results of previous studies. By using theoretical models and frameworks of assessment, instruments were developed to assess performance in practice and the progress of performance over a prolonged period. Now we have a programmatic framework for assessment that can be employed from the beginning until the end of the training. The instruments proved to be reliable and valid. A major finding is that the validated Compass, using the standard upon completion of the three-year training programme as a fixed reference point for judgment, demonstrates progress of performance in all the competencies of GP-trainees.

It seems that the non-medical competencies, Collaboration, Management, Organization and Health Advocacy, are more difficult to assess than the medical competencies Medical Expertise, Communication and Professionalism. However, these non-medical competencies will become increasingly relevant in primary care in view of ongoing changes in society and health care. Currently, no procedures are available to assess these competencies and must therefore be developed. The practice-based assignments that are being developed may be an appropriate answer.

In the development and evaluation of competency-based assessment procedures, we recommend to make use of the trainers' expertise, as this will enable the instruments to be better tailored to use in daily practice and to be more relevant to education.
References

Chapter 6


Samenvatting
Samenvatting

Dit proefschrift richt zich op de ontwikkeling en validering van competentiegerichte procedures waarmee het functioneren van de arts in opleiding tot specialist (aios) op de werkplek kan worden beoordeeld. Deze procedures omvatten een breed scala van competenties en laten de groei van de aios zien. De ontwikkeling van deze procedures was gebaseerd op theoretische modellen en raamwerken van toetsing, met name Miller’s indeling van bekwaamheid, raamwerken van toetsing op de werkplek, zoals voorgesteld door Pangaro en Ten Cate, en het conceptuele model van programmatische toetsing ontwikkeld door Van der Vleuten.

In het onderzoek hebben we ons gericht op de inhoud, betrouwbaarheid, haalbaarheid en op het proces. We voerden het onderzoek uit in de 'dagelijkse wereld' van de huisartsopleiding en waren daarom genoodzaakt om onze studie aan te passen aan de implementatie van de doorlopende ontwikkelingen op het gebied van het medisch onderwijs. In dit hoofdstuk worden achtereenvolgens besproken: de belangrijkste resultaten, hoe deze bevindingen te interpreteren in relatie tot de literatuur, de methodologische overwegingen, wat deze studie toevoge, vervolgonderzoek, praktische aanbevelingen en de conclusie.

Belangrijkste resultaten
De belangrijkste bevinding van onze studie is dat het met de Competentie Beoordelings Lijst (ComBeL) mogelijk is de groei van alle competenties van de aios te laten zien. De resultaten onderstrepen het belang om bij de beoordeling van competenties een vast referentiepunt te gebruiken. In de ComBeL die wij onderzochten, was het referentiepunt voor de beoordeling het eindniveau van de huisartsopleiding. Met dit referentiepunt was het mogelijk de voortgang in de tijd zichtbaar te maken. In de volgende secties worden de resultaten van de vier studies gepresenteerd.

Nijmegen Professionalisme Schaal
De Nijmegen Professionalisme Schaal is een instrument waarmee het professioneel gedrag van de aios huisartsengeneeskundige getoetst wordt. De lijst bestaat uit vier thema's: professioneel gedrag ten aanzien van patiënten, ten aanzien van collega's, ten aanzien van de maatschappij en ten aanzien van de aios zelf. Om de constructvaliditeit van de Nijmegen Professionalisme Schaal te onderzoeken hebben we een factoranalyse uitgevoerd. Daaruit bleek dat de lijst eenvoudiger kan worden opgezet, met twee dimensies voor elk thema. In het thema "professioneel gedrag ten aanzien van patiënten" hebben we de factoren respect voor patiëntenbelangen en professionele afstand genoemd. In het thema "professioneel gedrag ten aanzien van collega's" noemden we de factoren samenwerken en als leidinggeven, in het derde thema "professioneel gedrag ten aanzien van de maatschappij" verantwoordelijkheid en kwaliteit. Tenslotte in het vierde thema "professioneel gedrag ten aanzien van de aios zelf" reflectie & leren en omgaan met emoties. Vervolgens hebben we de factorstructuur van de lijsten die waren ingevuld door de
Samenvatting

opleiders vergeleken met die van de zelfevaluatie van de aios. We vonden geen grote verschillen tussen beide.

Cronbach's alfa coëfficiënten varieerden van 0.79 (omgaan met emoties) tot 0.95 (reflectie en leren), dus de interne consistente binnen elk thema is goed te noemen. Gegevens waaruit zou blijken dat de Nijmegen Professionalisme Schaal in staat is om huisartsen in opleiding met lacunes op het gebied van professioneel gedrag te kunnen identificeren zou de validiteit van het instrument nog meer ondersteunen. Verder zou het instrument ook de voortgang van Professionaliteit van de aios moeten kunnen meten. In principe zou dit mogelijk moeten zijn omdat we geen plafondefecten hebben geconstateerd, maar in onze studie naar de Nijmegen Professionalisme Schaal hebben we dit niet verder onderzocht omdat op dat tijdstip de Competentie Beoordelings Lijst (ComBeL) werd ingevoerd in de huisartsopleiding. De ComBeL bevat niet alleen elementen van de Nijmegen Professionalisme Schaal maar ook het format van longitudinaal toetsen.

De ComBeL
De ComBeL is een beoordelingslijst waarmee op programmatische wijze elke drie maanden een beoordeling gegeven wordt over de zeven competenties die aios in de huisartsopleiding moeten verwerven. Het is de bedoeling dat de opleider elke drie maanden de ComBel invult gebruikmakend van verschillende instrumenten (bijvoorbeeld de KKB) en bronnen (bijvoorbeeld feedback van praktijkassistentes) tijdens verschillende praktijksituaties. De huisartsopleiders aggregeren de resultaten in de ComBeL en gebruiken deze om feedback te geven tijdens de driemaandelijkse voortgangsgesprekken.

De oorspronkelijke versie van de ComBeL was ontwikkeld door ervaren huisartsen en docenten betrokken bij de huisartsopleiding. Uitgangspunt waren de CanMEDS competenties, aangepast aan de Nederlandse huisartsgeneeskunde. Uit onze validatiestudie, waarin we de RAND Delphi consensus methode gebruikt hebben, bleek dat maar enkele kleine inhoudelijke veranderingen nodig waren. Alle schalen vertoonden een goede interne consistentie met waarden die varieerden van .89 tot .94.

Met de ComBeL is het mogelijk de groei van alle competenties van de aios te laten zien. De gemiddelde scores van de aios werden hoger naarmate de opleiding vorderde.

Om inzicht te krijgen hoe opleiders tot hun oordeel komen, hebben we een kwalitatieve studie uitgevoerd. Opleiders geven in de ComBeL een toelichting op hun oordeel door in het vakje 'overwegingen' te schrijven welke factoren een rol hebben gespeeld bij het bepalen van de scores. De inhoud van deze overwegingen hebben we g-analyseerd. Uit het onderzoek bleek dat er twee categorieën competenties konden worden onderscheiden. De competenties Medisch
Samenvatting

Handelen, Communicatie en Professionaliteit vormden de eerste categorie, de tweede categorie bestond uit Samenwerken, Organisatie, Maatschappelijk Handelen en Wetenschap & Onderwijs. De overwegingen uit de eerste categorie waren uitgebreider en specifieker en waren gericht op hoe de aios functioneert als dokter, op haar/zijn ontwikkeling en persoonlijke eigenschappen. In de tweede categorie kwamen 'voorlopige' beoordelingen voor. Daarnaast vonden we dat in deze categorie de beoordelingen per competentie niet scherp afgebakend waren: bij de beoordeling van één competentie speelden ook andere competenties een belangrijke rol.

Selectieprocedure
We hebben een selectieprocedure ontwikkeld gebaseerd op empirische en theoretische studies. Om de kwaliteit van de selectieprocedure te waarborgen, beschouwden we selectie als "toetsing voor selectie". De eerste stap in de ontwikkeling was een inhoudsanalyse, waarbij we van de CanMEDS competenties uitgingen. De volgende competenties zijn uiteindelijk geïncludeerd in de procedure: Medisch Handelen, Communicatie, Samenwerken, Organisatie en Professionaliteit. Vervolgens hebben we besloten om diverse instrumenten te gebruiken die informatie konden bieden over de verschillende elementen van de te toetsen competenties. Geleid door de literatuur maar ook door haalbaarheid is gekozen voor een viertal instrumenten: een kennisptoets, een situationele beoordelingstest (deze test geeft een beeld hoe kandidaten situaties inschatten waarmee ze worden geconfronteerd in een professionele context), een serie met drie praktijk gerelateerde simulaties en een competentiegericht interview. Omdat de inhoud van de selectieprocedure overeenkomt met de inhoud van het curriculum, kunnen zowel aios als opleiders de resultaten van de procedure gebruiken om te bepalen waar zich tekortkomingen bevinden. Zo kunnen leerplannen worden geformuleerd in het vroegst denkbare stadium van de opleiding.

Interpretatie van de bevindingen in relatie tot de literatuur
Om de constructvaliditeit van de Nijmegen Professionalisme Schaal te onderzoeken hebben we een factoranalyse uitgevoerd. De oorspronkelijke structuur was gebaseerd op consensus van experts op het gebied van professionalisme. Onze factoranalyse liet een veel eenvoudigere structuur zien met twee dimensies voor elk thema. Naar onze mening maakt deze nieuwe structuur het instrument conceptueel duidelijker en daardoor voor de opleiders makkelijker te gebruiken.

Vervolgens hebben we de factorstructuur van de lijsten die waren ingevuld door de opleiders vergeleken met die van de zelfevaluaties van de aios. We vonden geen grote verschillen tussen beide. Dit betekent dat aios en huisartsopleiders op dezelfde wijze invulling geven aan het construct "professioneel gedrag". Deze bevinding verhoogt de bruikbaarheid van het instrument, omdat het oordeel en de feedback van de opleider aansluiten bij de ervaringen van de aios. We
Samenvatting

hebben de structuur van de Nijmegen Professionalisme Schaal aangepast op basis van de resultaten van de factoranalyses.

De nieuwe, eenvoudige lijst, sluit beter aan bij hoe de opleiders naar professionaliteit kijken. Dat komt de validiteit en betrouwbaarheid ten goede. Het onderzoek laat zien dat bij de ontwikkeling van een beoordelingsinstrument ook de uiteindelijke gebruikers van een instrument betrokken moeten worden. Crossley en Jolly kwamen tot dezelfde conclusie: ook zij benadruktten dat er zoveel mogelijk gebruik moet worden gemaakt van de expertise van de gebruikers van een te ontwikkelen instrument. Opleiders in een keurslijf van een opgelegde inhoud en structuur dwingen komt volgens deze onderzoekers de beoordeling van de competency's van aios niet ten goede.

In de ontwikkeling van ComBeL, kozen we doelbewust voor een vast referentiepunt. Onze resultaten lieten zien dat het vaste referentiepunt voor de beoordeling, in dit geval het eindniveau van de opleiding, opleiders duidelijkheid verschaf hoe het handelen op de werkplek kan worden beoordeeld door het naast een vooraf bepaalde, duidelijke standaard te leggen. Om aios en opleiders duidelijkheid te verschaffen wat er in de praktijk geëvalueerd moet worden, welke observaties of beoordelingen daarbij bruikbaar kunnen zijn en hoe het handelen op de werkplek kan worden beoordeeld door het naast een vooraf bepaalde standaard te leggen, zijn raamwerken van beoordeling op de werkplek nodig. Bij iedere formele beoordeling op de werkplek hoort een expliciete maatstaf, een referentiepunt om te kunnen bepalen of een aios het vereiste niveau al dan niet heeft bereikt. Met observatie-instrumenten zijn er vaak problemen wat betreft het referentiepunt waartegen de prestaties van de aios moet worden afgezet. Vaak blijft dit referentiepunt impliciet en wordt het overgelaten aan de beoordelaars. Prescott-Clements en haar collega's stellen dat toetsinstrumenten die niet het eindpunt van de opleiding als referentiepunt nemen, weinig variatie vertonen gedurende de opleiding. Deze onderzoekers ontwikkelden een instrument, "the longitudinally evaluated performance" (LEP), waarmee longitudinaal de prestaties tijdens de opleiding worden beoordeeld. Zij rapporteerden dat veel opleiders het gebruik van dit vaste referentiepunt voor de beoordeling ondersteunden, terwijl er vooraf bij die opleiders veel twijfels leefden. Meer dan zeventig procent van de opleiders was van mening dat deze werkwijze bijdroeg aan de ontwikkeling van de aios tijdens de opleiding en meer dan negentig procent dacht dat op deze manier groei tijdens de opleiding zichtbaar gemaakt kon worden. Uit eerder onderzoek bleek dat het moeilijk was om groei in competency's aan te tonen. Kramer en haar collega's bijvoorbeeld, slaagden er niet in om de groei van de communicatievaardigheden van aios tijdens de huisartsopleiding aan te tonen, hoewel de aios aan het eind van de opleiding aangaven dat ze zich zekerder voelden in vergelijking met het begin van de opleiding.
In een kwalitatieve studie vonden we dat de overwegingen van opleiders hoe zij tot hun oordeel kwamen uitgebreider werden opgeschreven voor de competenties Medisch Handelen, Communicatie en Professionaliteit. Beoordeling van de competenties Samenwerken, Organisatie, Maatschappelijk Handelen en Wetenschap & Onderwijs leek iets meer problemen op te leveren. Mogelijk kennen opleiders ze zelf nog onvoldoende of hebben ze onvoldoende mogelijkheid om ze te beoordelen. Deze nieuwe competenties beperken zich immers niet alleen tot de spreekkamer. Voor de beoordeling van de meer traditionele competenties Medisch Handelen, Communicatie en Professionaliteit zijn verschillende instrumenten beschikbaar, zoals de Maas Globaal en de Nijmegen Professionalisme Schaal. Voor de nieuwe competenties moeten nieuwe procedures ontwikkeld worden om de volle breedte van het competentiegerichte onderwijs te kunnen benutten. De nieuwe competenties verwijzen naar vaardigheden die in de toekomst nodig zijn ten gevolge van de huidige ontwikkelingen in de gezondheidszorg en daarmee in de huisartsengeneeskunde. In een enquête van het NIVEL gaven aios aan dat er, volgens hen, tijdens de opleiding te weinig aandacht werd besteed aan de competenties Organisatie en Maatschappelijk Handelen. Onze resultaten benadrukken dat er meer aandacht moet worden besteed aan de nieuwe competenties. In de Nederlandse huisartsopleiding heeft dit geleid tot de ontwikkeling van Competentie Gerichte Opdrachten (CGO). Deze opdrachten weerspiegelen de complexe dagelijkse praktijk, waar competenties niet in isolatie voorkomen maar met elkaar vervlochten zijn. Met CGO’s kunnen meerdere competenties tegelijkertijd beoordeeld worden. Deze benadering van CGO’s is vergelijkbaar met het concept van "entrustable professional activities (EPA’s)" dat ontwikkeld is door ten Cate en collega’s. Deze auteurs zijn van mening dat het scheiden van de zeven competenties kunstmatig is, in plaats daarvan zou men competenties moeten onderwijzen en toetsen op een geïntegreerde manier. De invoering van CGO’s is veelbelovend, maar de bruikbaarheid voor de opleiding moet nog onderzocht worden.

Competentiegericht onderwijs is gericht op individueel en zelfsturend leren. Daarom is het van belang om het niveau van de aios in het begin van de opleiding te bepalen. Beginnende aios vormen namelijk een heterogene groep. Zij hebben een verschillend competentieniveau als gevolg van verschillende werkvaringen. Regelmatig spreken opleiders de zorg uit dat een gedeelte van de afgestudeerde basisartsen niet voldoende is toegerust om aan een medische vervolgopleiding te kunnen beginnen. Deze situatie is niet uniek voor Nederland. In Denemarken bijvoorbeeld, gaven opleiders aan dat gemiddeld 25% van de beginnende aios niet de klinische vaardigheden bezitten die verwacht mogen worden na de basisopleiding.

Een selectieprocedure kan worden beschouwd als een beginnend om het niveau van de aios te bepalen. Om de kwaliteit van de huidige selectieprocedure te verbeteren, beschouwden we selectie als "toetsing voor selectie". Zo konden we gebruik maken van de "best practice" principes van competentiegericht toetsen. Een van deze principes is dat het gebruik van één instrument niet toereikend is om daar een belangrijke beslissing, zoals selectie, van af te laten
hangen. Geaccepteerd of afgewezen worden voor een opleiding is een zwaarwegende beslissing, daarom zijn er meerdere instrumenten in de procedure opgenomen.

Een ander "best practice" principe is dat toetsing moet zijn gebaseerd op een inhoudsanalyse, er moet een blauwdruk zijn van inhoud en de toetsing moet zich ook tot die inhoud beperken. Zover wij weten is dit de eerste selectieprocedure die gebaseerd is op de CanMEDS competenties. Kenmerkend voor deze procedure is dat we vooraf bepaald hebben welke competenties tijdens de procedure getoetst moesten worden. Het CanMEDS-raamwerk geeft inzicht wat een arts moet kennen en kunnen. Sommige competenties echter, zoals bijvoorbeeld het tonen van empatie of het vermogen om te reflecteren over het eigen gedrag, zijn zeer moeilijk aan te leren, en zouden al ontwikkeld moeten zijn gedurende de basisopleiding. Andere competenties, zoals bijvoorbeeld kleine chirurgische ingrepen, kunnen makkelijker worden aangeleerd tijdens de vervolgsopleiding. Dit onderscheid is essentieel bij de bepaling van de selectiecriteria. Onze studie maakt duidelijk dat met behulp van deze overwegingen een panel van ervaringsdeskundigen op het gebied van selectie en van de huisartsopleiding kan bepalen welke CanMEDS-competenties relevant zijn voor de selectieprocedure. Door de competenties van ons curriculum als basis voor de selectieprocedure te nemen, en door diezelfde competenties te toetsen die ook tijdens de opleiding worden getoetst, hebben we de selectie in lijn gebracht met het toetsprogramma van de opleiding. Daardoor kunnen we de selectie als een eerste toetsing van dit toetsprogramma beschouwen. In het competentiegericht toetsen staat het begeleiden van de professionele ontwikkeling van de arts in opleiding centraal. Feedback is daar een belangrijk onderdeel van. Ook in de selectieprocedure speelt feedback een rol. Opleiders en startende aios kunnen de resultaten van de selectie-instrumenten gebruiken om te bepalen waar zich tekortkomingen bevinden en dit gebruiken om leerplannen te formuleren. Zo kunnen mogelijke zwakke punten al in het vroegste stadium van de opleiding onder de aandacht worden gebracht en opgenomen worden in de eerste leerplannen. De mogelijkheid om zo vroeg in de opleiding al feedback te kunnen geven en krijgen is uniek. Daarom kan deze selectieprocedure zowel voor de aios als voor de opleiding grote voordelen opleveren.

Methodologische overwegingen

Dit proefschrift beschrijft een aantal studies die als centrale vraag hebben: hoe werkt de nieuwe benadering van competentiegericht toetsen in de dagelijkse praktijk? We bestudeerden de validiteit van nieuw ontwikkelde instrumenten. In het bijzonder onderzochten we de psychometrische eigenschappen van instrumenten zoals die gebruikt worden tijdens de opleiding. Deze benadering heeft als voordeel dat bewijs wordt geleverd over het gebruik in de dagelijkse praktijk. Dit verhoogt de relevantie van de resultaten en verhoogt de relevantie voor het medisch onderwijs.
Samenvatting

Het aanwenden van een kwalitatieve onderzoeksmethode leverde informatie op die we niet hadden gevonden als we ons beperkt hadden tot alleen kwantitatieve methoden. Juist de resultaten van onze kwalitatieve studie gaven ons meer inzicht in hoe opleiders de ComBeL invullen. Deze resultaten roepen echter ook weer nieuwe vragen op, zoals hieronder zal worden besproken.

Een beperking van dit onderzoek is dat de studies zijn uitgevoerd in een van de acht huisartsopleidingen in Nederland. Daarom is het niet helemaal zeker of de resultaten generaliseerbaar zijn. Bovendien, hoewel de studies algemene thema’s van medisch onderwijs behandelen, is in de huisartsopleiding sprake van een bijzondere situatie waar één aios en één opleider gedurende een jaar samenwerken. Dit is een groot verschil vergeleken met andere specialismen.

Een mogelijk nadeel van voortgangsbeoordelingen is dat de opleider anticipeert op het feit dat scores beter moeten worden naarmate de opleiding vordert en dat hij/zij dus bij elke driemaandelijkse beoordeling een hoger cijfer geeft. Wat hier tegen pleit is dat we zagen dat bij de beoordeling van een aantal aios dezelfde cijfers werden gegeven in de loop van de opleiding. Een klein aantal werd zelfs lager beoordeeld dan in de vorige periode.

Tenslotte, we hebben in ons onderzoek alleen data verzameld van het eerste jaar van de opleiding, we weten niet of de ComBeL gevoelig genoeg is om ook in het tweede en derde jaar voortgang te kunnen meten.

Wat deze studie toevoegt

We hebben competentietegerichte procedures bestudeerd waarmee longitudinaal de voortgang van de aios kan worden beoordeeld. De belangrijkste bevinding van onze studie is dat het met de ComBeL mogelijk is de groei van alle competenties van de aios te laten zien. Het referentiepunt voor de beoordeling was het eindniveau van de huisartsopleiding. Met dit referentiepunt was het mogelijk de voortgang in de tijd zichtbaar te maken. Het accent van onderzoek naar toetsing in medisch onderwijs ligt op afzonderlijke instrumenten die competenties meten tijdens een enkel consult. Met de ComBeL worden de resultaten van toetsen die afgenomen zijn gedurende een periode van drie maanden geaggregeerd. Deze continue toetsing blijkt gevoelig genoeg te zijn om de voortgang van de aios te kunnen meten. Door de selectieprocedure in het toetsprogramma van de opleiding te integreren, wordt formatieve toetsing mogelijk vanaf de eerste dag van de opleiding. Zo kunnen mogelijke lacunes al in het vroegste stadium van de opleiding onder de aandacht worden gebracht en opgenomen worden in de eerste leerplannen.
Vervolgonderzoek

In hoofdstuk 5 beschreven we de ontwikkeling van een competentiegerichte selectieprocedure. De volgende stap is de evaluatie van de psychometrische eigenschappen van deze procedure. Juist omdat er voor de kandidaten veel van afhankt is het belangrijk beslissingen te nemen die gebaseerd zijn op een betrouwbare en valide toetsing.

In hoofdstuk 4 vonden we dat opleiders niet altijd opschreven welke bronnen en instrumenten een belangrijke rol speelden in de beoordeling met de ComBeL. Verder onderzoek kan verduidelijken welke bronnen en instrumenten opleiders in de beoordeling betrekken, hoe zij informatie verzamelen en hoe zij die informatie gebruiken om tot een oordeel te komen. Ook van belang is hoe opleiders hun verzamelde informatie documenteren, maken zij gebruik van een logboek of plaatsen zij de informatie in een dossier?

Feedback speelt een belangrijke rol in competentiegericht medisch onderwijs. Voor de competenties Organisatie, Samenwerken, Maatschappelijk Handelen en Wetenschap & Onderwijs lijkt het dat opleiders onvoldoende waarnemen om goede feedback te kunnen geven. Er is onderzoek nodig om inzicht te krijgen in hoe de competenties het beste geobserveerd en beoordeeld kunnen worden. Alleen de ontwikkeling van instrumenten is niet genoeg. De gebruikers van toetsinstrumenten spelen een belangrijke rol. Daarom zou in de ontwikkeling ook een training moeten worden betrokken om de gebruikers leren om te gaan met de nieuwe instrumenten. Tot slot zou het effect van de nieuw ontwikkelde toetsprocedure onderzocht moeten worden. Worden er bijvoorbeeld meer leerdoelen over deze competenties geformuleerd, wordt er meer of specifieker feedback gegeven?

We hebben niet onderzocht of de ComBeL in het tweede en derde jaar in staat is om groei te meten. Misschien vermindert de voortgang in die jaren. Uit onderzoek blijkt dat medische kennis in het eerste jaar toeneemt, stabiliseert in het tweede jaar en weer iets toeneemt in het derde jaar. Het verdient aanbeveling om te onderzoeken hoe de ComBeL zich in het tweede en derde jaar gedraagt.

Praktische aanbevelingen

Op grond van onze bevindingen adviseren wij om bij voortgangsbeoordelingen het eindniveau van de opleiding als referentiepunt te gebruiken. Dit referentiepunt geeft een duidelijk beeld voor zowel aios als opleider. Vaak blijft dit vaak impliciet en wordt het overgelaten aan de beoordelaars.

Het competentiegericht medisch onderwijs kan beschouwd worden als 'werk in uitvoering' daarom zullen nog verscheidene veranderingen volgen. Een van de voorwaarden om veranderingen succesvol te implementeren is dat leidinggevenden consensus bereiken omtrent
veranderingen en deze faciliteren, vooral als deze gebaseerd zijn op wetenschappelijk bewijs. De resultaten van de studie over de ComBeL zijn in 2012 gepubliceerd. Toch is een richtlijn uitgevaardigd voor het gebruik van de ComBeL waarin referentiepunt de fase van de opleiding is in plaats van het eindniveau.

Het is aan te bevelen opleiders te betrekken in de ontwikkeling van instrumenten. Zij zijn tenslotte bekend met de dagelijkse gang van zaken in de praktijk en hun betrokkenheid kan de validiteit en betrouwbaarheid, en ook de implementatie verbeteren.

Wij bevelen aan om selectie, curriculum en toetsing op één lijn te brengen, door tijdens de selectie dezelfde competenties te toetsen als tijdens de opleiding. Daarmee wordt de selectieprocedure de eerste toetsing in een toetsprogramma. Ook andere specialismen waarvan het curriculum gebaseerd is op de CanMEDS of een ander competentieraamwerk, kunnen dit model gebruiken.

Conclusie
In dit proefschrift hebben we ons gericht op de ontwikkeling en validering van competentiegerichte procedures waarmee het functioneren van de arts in opleiding tot specialist (aios) op de werkplek kan worden beoordeeld. De ontwikkeling van de instrumenten waarmee de groei van het handelen in de praktijk kan worden beoordeeld was gebaseerd op theoretische modellen en raamwerken van toetsing. We beschikken nu over een programmatisch raamwerk van toetsing dat te gebruiken is vanaf de eerste dag van de opleiding. De instrumenten die we onderzochten bleken betrouwbaar en valide te zijn.

Een belangrijke bevinding is dat de ComBeL, met als referentiepunt voor de beoordeling het eindniveau van de huisartsopleiding, voortgang laat zien van alle competenties van aios.

Het lijkt erop dat de competenties Samenwerken, Organiseren, Maatschappelijk Handelen en Wetenschap & Onderwijs moeilijker te beoordelen zijn dan de competenties Medisch Handelen, Communicatie en Professionaliteit. De "niet-medische" competenties zullen echter in de toekomst steeds belangrijker worden, gezien de ontwikkelingen in de maatschappij en in de gezondheidszorg. Tot op heden zijn nog geen procedures beschikbaar om ze te beoordelen. Deze moeten worden ontwikkeld. De competentiegerichte opdrachten die nu worden ontwikkeld zouden een passende oplossing kunnen zijn. In de ontwikkeling en evaluatie van deze procedures is het raadzaam om van de expertise van opleiders gebruik te maken. Op deze manier zullen ze geschikter zijn voor gebruik in de praktijk en dus relevanter voor de opleiding.
Dankwoord
Dankwoord

Als ik terugkijk op de afgelopen jaren waarin ik me bezig gehouden heb met dit onderzoek, besef ik dat ik de samenwerking met alle betrokkenen niet had willen missen. Nu wil ik iedereen bedanken zonder wie dit proefschrift nooit tot stand had kunnen komen. Het werken aan dit proefschrift was niet altijd even gemakkelijk. In 2008 kreeg ik gezondheidsproblemen en moest ik me opnieuw bezinnen op mijn toekomst. Ik heb daar alle ruimte voor gekregen en daar ben ik iedereen erg dankbaar voor. De ziekte had gelukkig een rustig beloop en ik kon me weer snel op het onderzoek richten.

Allereerst wil ik mijn promotor en copromotoren bedanken. Myrra, dank voor je vertrouwen en geduld. Met dit proefschrift eindigt een lange periode van begeleiding. Ik heb veel van je geleerd. Je moest me vaak weer "bij de les brengen". Je wist me altijd weer in de goede richting te krijgen met je vakkundige en vaak ook pittige feedback. Toch heb je me de ruimte gegeven om me te kunnen ontwikkelen. Nu ligt het proefschrift er dan toch. Zonder jou was het niet gelukt.

Ben, jij bood mij de gelegenheid om te gaan werken aan de ontwikkeling van een selectieprocedure voor de huisartsopleiding. We hebben daarin intensief samengewerkt. Jouw uitgebreide vakkennis, ideeën, suggesties en commentaren waren inspirerend en hebben er voor gezorgd dat er nu een competentiegerichte selectieprocedure beschikbaar is.

Anneke, je bent altijd goedgehumeurd en het is een plezier om met je te werken. Je kennis over medisch onderwijs was onmisbaar en je wist altijd wel een oplossing te vinden als ik even niet meer verder kwam met een artikel. Ook gaf me de ruimte om te kunnen ontwikkelen. Hartelijk dank voor je goede inbreng.

Richard, je bent nu met pensioen, maar toen ik begon aan dit project heb je me begeleid. Je was altijd een stap voor op alle ontwikkelingen in het medisch onderwijs en mede daardoor heeft het proefschrift zijn uiteindelijke vorm gekregen.

Ook heb ik veel steun gehad van al mijn collega's.

Bart, toen ik begon bij de VOHA was je mijn eerste kamergenoot. Ik voelde me meteen thuis door je humor en je gezelligheid. Je hebt me goed op weg geholpen door me de ins en outs van de huisartsopleiding te leren kennen.

Henk, ik heb veel van je geleerd, heel veel dank voor je hulp en ondersteuning bij ingewikkelde analyses. We hebben veel gepraat, niet alleen over onderzoek maar ook over persoonlijke dingen. Het was fijn om je te leren kennen.
Els, Greetje, Geurt, Patrick, jullie waren fijne collega's. De verschillende congressen die we met zijn allen hebben bezocht, de wandelingen, de gesprekken, de adviezen, ik had het niet willen missen.

Mijn paranimfen
Thea, we ontmoetten elkaar tijdens de cursus kwalitatief onderzoek. Kort daarna kwam je in Nijmegen werken. Onze samenwerking heeft twee artikelen opgeleverd. Vanaf de eerste dag konden we het goed met elkaar vinden, ik was altijd welkom voor een advies, of een zo maar een praatje. Je werkt nu in Eindhoven en ik mis onze samenwerking, gesprekken, wandelingen en grappen nog elke dag.

Margit, we deden alle twee onderzoek naar selectie voor medische vervolgopleidingen, zo is onze samenwerking begonnen. Het was een groot project en dat gaat met de nodige ups en downs. Ik heb van samenwerking genoten, je bent een prettige collega. Ook heb ik veel van je geleerd. Je bent niet snel tevreden en altijd op zoek om nog betere resultaten te krijgen. Dat heb ik erg gewaardeerd.

Tenslotte de allerbelangrijkste in mijn leven.
Lieve Susanne, we delen al dertig jaar lief en leed met elkaar. Je hebt geduldig mijn verhalen over toetsen, competenties, medisch onderwijs en selectie steeds weer aangehoord. Je hebt altijd in mij geloofd en je hebt me alle ruimte gegeven om dit te kunnen doen. Ik heb erg veel geluk gehad dat ik jou ben tegengekomen, het was in 1982 in theater "De Bergruimte" te Eindhoven. Ik hoop dat we samen nog veel mooie momenten mogen meemaken.