Performance assessment: climbing the pyramid

Recent developments in assessment in health professions education are characterized by an increasing emphasis on assessment of performance, targeting the upper levels of Miller’s pyramid. Reasons for the growing emphasis on what students do or can do, rather than simply on what they know or what they say they will do, are manifold (Chapter 1). Major drivers for change of assessment practices are the rise of competency- or outcome-based models of education and increasing public pressure for educational and professional accountability, creating a need for assessment methods that require integration of theoretical and practical knowledge in task performance, thereby providing more direct evidence of the proficiencies of interest, i.e. performance in practice, or readiness for advancement in training (Kane, 2006; Kane et al., 1999).

Furthermore, changing conceptions of what constitutes ‘good medical practice’ or ‘a good physician’, as well as rapidly changing working environments demand new competencies which cannot be validly assessed with traditional, cognition-based ‘off-the-job’ assessments: teamwork, professionalism, interpersonal skills, practice-based learning and life-long learning are considered key competencies of today’s health care professionals.

Finally, there is conclusive evidence that prolonged task experience is not enough to ensure professional expertise: competence is not a ‘once-in-a-lifetime’ achievement (e.g. Choudhry et al., 2005). Professionals must construct and reconstruct their expertise throughout their professional careers, in a process of lifelong learning. Evaluation of and feedback on performance are critical variables for performance improvement and maintenance of professional competence (Ericsson, 2004; 2009).

Therefore, performance assessments fulfilling summative and formative assessment functions will become a sine qua non in assessment programmes in educational contexts, but they will also become part of health care workers’ professional life.

Figure 8.1 Miller’s pyramid (adapted from: Miller, 1990)
Understanding performance assessment

Although the shift in focus towards performance assessments is undisputed, concerns about the utility of assessments and assessment outcomes may limit their use in educational and professional settings (Chapter 1). The studies presented in this thesis addressed several key issues in performance assessment, aiming to further our understanding of how to improve utility of assessment and assessment outcomes in the evaluation of professional competence. Research presented in this thesis reflects developments in competence assessment as described in chapter 1. Climbing Miller’s pyramid, our first studies of standardized performance-based assessment in midwifery education were followed by studies of assessment in ‘real-life’ settings, both in midwifery and medical education.

More specifically, our research questions addressed:
- the utility of performance-based tests in relation to the task domain and ways to improve test efficiency;
- the utility of a systematic and semi-standardized approach towards WBA in clinical rotations;
- processes underlying judgment and decision making in performance assessments in the clinical context, focusing on:
  - factors that influence assessment outcomes in real-life settings;
  - raters’ cognitive processing when observing and judging trainee performance in patient encounters.

Utility of standardized performance assessment: effects of task domain, rating scale and score interpretation

In chapters 2 and 3 we describe the development, implementation, and evaluation of a standardized performance-based test to assess the professional competence of (future) midwives. Based on research findings in medical education and an integrated conception of professional competence, a 6-station, 3-hour OSCE-based examination was developed, focusing on risk assessment and patient management in midwifery. In our studies we addressed several aspects of assessment utility. We used a survey to examine both students’ and examiners’ perceptions of test acceptability and educational consequences. Additional studies addressed the generalizability of test results and the possibilities for increasing test efficiency.

In general, findings within the midwifery domain are in line with findings in the medical domain (Reznick and Rajaratnam, 2000; Howley, 2004; Yudkowsky, 2009).

As was to be expected, we found high acceptability of the performance-based test for the assessment of professional competence. Both students and examiners felt that the assessment of student performance in handling real-life professional tasks, focusing on students’ ability to integrate knowledge, skills and professional judgment in task performance, contributed to the authenticity, meaningfulness and fairness of the assessment programme in midwifery education. More importantly, however, our findings indicate that performance-based assessments may have a significant educational impact, confirming research findings in other domains.
(Struyven et al., 2003; McDowell, 1995). Implementation of the more ‘authentic’ assessments resulted in positive effects on both student learning, teaching methods and curriculum content – through fostering reflection and discussion on key competencies and ‘good medical practice’ in midwifery care.

Findings, however, also highlight the complexity and resource-intensiveness of development and administration of performance-based tests. Extensive testing time (more than 20 cases or 10 hours of testing time) is needed to achieve minimum acceptable levels of reproducibility (0.80), even in a relatively narrow domain such as midwifery. Findings are consistent with findings in medical education (Colliver et al., 1991; Singer et al, 1996; Keller et al., 2010; Clauser et al., 2006). Our findings demonstrate that test efficiency can be enhanced by the use of global rating scales – rather than task- or case-specific checklists. Global ratings, which rely on expert judgments, are better able to capture expertise and are more effective in differentiating between examinees (Hodges et al., 1999; Regehr et al., 1998; Cunnington et al., 1997), thereby reducing test length requirements. Test efficiency can be enhanced even further by the adoption of a mastery-oriented perspective to score interpretation. The borderline regression method, as described in chapter 3, which combines global ratings of trainee competence with more detailed scoring of performance by a large panel of experts (i.e. examiners), provides a standard setting procedure that is relatively inexpensive as well as credible, defensible and easy to explain and implement (e.g. Kramer et al., 2003). These findings - perhaps counter intuitively- indicate that careful / judicious use of human expert judgment in standardized performance-based assessments may add significantly to assessment quality.

**Work-based assessments: effects of structuring – standardizing the real world**

Chapter 4 describes a study of the use of observational diaries in clerkship assessment in midwifery training. The use of observational diaries was based on principles of extensive work sampling and frequent (daily) documentation of feedback and performance evaluations through the use of standardized, detailed checklists. The main purposes of the observational diaries were to support and guide student learning throughout the clerkship. Documentation of formative feedback and performance evaluations in the diary was also used qualitatively to support decision making about competence development and achievement. Frequent and continuous performance assessments were thus used for both formative and summative assessment purposes.

Findings from our study clearly indicate that formalizing assessment in work settings may enhance student (or professional) learning. Frequent documentation of feedback and evaluation of performance against preset criteria and performance standards support competence development through helping students and supervisors focus on essential learning goals and competencies that might otherwise be neglected. More importantly, adequate and adequately documented feedback gives insight into progression in relevant competencies, thereby increasing students’ self-confidence in their ability to achieve end-of-clerkship educational goals. Observational diaries or, more specifically, regular documentation of feedback and performance
evaluations may therefore fulfil a crucial role in making work-based learning more explicit (Teunissen, 2009). Learning is enhanced through engaging students in the assessment process. Discussion of and guided reflection on feedback as well as incorporation of self-assessments are central to students’ perceptions of whether work-based assessment is meaningful. These findings are in line with recent findings in medical education (Watling et al., 2008) and in other domains (Taylor et al., 1995; Erdogan et al., 2001; Topping, 2003).

Findings from our study furthermore indicate that rating formats may have variable utility in steering learning, depending on students’ level of competence. Standardization of performance in detailed checklists rapidly becomes less important - and even meaningless - as expertise increases. In general, narrative feedback as well as ‘analytic global rating scales’ which can be attuned to individual learning and expertise development are far more effective in guiding student learning than analytic numerical scoring. Narrative feedback provides meaningful information about strengths and weaknesses and enhances reflection on performance through explicit linkage of performance effectiveness with task-specific factors, putting behaviours in context (Korthagen and Vasalos, 2005).

As expected, students feel that ‘real-life’ work-based assessments provide the most valid way of assessing professional competence, i.e. habitual performance (Epstein and Hundert, 2002). Clearly, the use of observational diaries enhances the utility of work-based assessments through promoting direct observation, evaluation and documentation of performance. Findings from our study, however, also confirm research findings with respect to the use of expert judgments (e.g. Albanese, 2000). More specifically, rater idiosyncrasy is perceived as one of the major impediments to using performance ratings for summative purposes. Meaningful interpretation of performance scores therefore requires additional narrative feedback and comments providing insight into the rater’s personal motivations and arguments underlying judgment and decision making. Consequently, adequate descriptions of observed behaviours relating to both performance theory and task-specific factors affecting performance effectiveness are considered at least as important and informative for summative decision making as numerical scores.

Our results indicate that formative and summative functions can and should be combined in work-based assessment (WBA). Summative use of assessment can enhance its formative function by requiring raters to regularly observe task performance, provide feedback and carefully document performance evaluations. Frequent formative performance evaluations, on the other hand, are prerequisites for credible and defensible (valid) decision making. Optimizing the utility of WBA for formative as well as summative purposes therefore requires a delicate balance between frequent documentation of non-judgmental feedback and regular, but less frequent, normative judgments about progress and competence achievement. In this respect, our findings support current assessment trends in competency-based medical education and residency training (Ten Cate, 2007).

Finally, our findings confirm the importance of the learning climate for useful and effective WBA. Psychological safety, openness, transparency of assessment, honesty and due process are key factors determining trust in and acceptance of the assessment system and preconditions for actual use of assessment outcomes for performance improvement (Sargeant et al., 2008).
Why work-based assessment may not always work

Chapters 2-4 underline the central role of ‘expert judgment’ or ‘professional judgment’ in performance assessment. The obvious need for and reliance on expert judgments, however, seems to be at odds with quite consistent research findings pointing out the fallibility of human judgment and decision making in behavioural domains (e.g. Shanteau, 1992). Chapter 5 makes significant contributions to a better understanding of judgment and decision making by professionals in assessment situations.

Drawing on research in various professional fields and based on insights from other disciplines, we propose an approach towards WBA which takes a predominantly constructivist, social-psychological perspective and integrates elements of theories of (social) cognition, motivation and decision making.

Central to our approach is that raters are no longer to be seen as passive measurement instruments, but as active information processors, who interpret and construct their own personal reality of the assessment context. We argue that expert judgments are inherently idiosyncratic. Raters will construct and reconstruct their own performance theories through training, socialization and relevant task experience. Consequently, multiple raters will have multiple constructed realities (Van der Vleuten et al., 2010). We herewith challenge the assumption of a ‘true score’ and the exclusive use of the psychometric framework in assessment research.

We furthermore propose that assessment of performance in real life settings is a judgment and decision making process in which raters’ behaviours are shaped by interactions between individuals and the social context in which the assessment occurs (Levy and Williams, 2004). We take the view that raters’ behaviours are motivated and goal-directed, defined by raters’ perceptions of the assessment system and its intended or unintended (negative) effects. Consequently, actual public ratings may communicate raters’ goals to other parties involved in the assessment process and actual ratings may differ from raters’ personal judgments or feedback to ratees (Murphy and Cleveland, 1995).

From the perspective of our approach, it does not make much sense to exclusively attribute raters’ errors to raters’ inability to produce accurate ratings. Discrepancies between actual performance and ratings may simply reflect effects of forces that discourage accurate rating, while failure to discriminate between persons or dimensions may constitute adaptive behaviour. In fact, raters’ behaviour may be driven more strongly by contextual variables than by actual differences between ratee variables (Murphy and Cleveland, 1995). Attempts to improve WBA in the health care domain might therefore need to focus more on factors within the (organizational) context in which performance assessments take place. Key factors appear to be organizational norms and values regarding performance assessment, transparency of assessment purposes and assessment process (due process), support, accountability and feelings of ‘psychological safety’ -based on open and honest communications and interactions between all stakeholders in the assessment process.
Focus on rater cognition

It is inherent in all performance assessments that all information has to pass through the cognitive filter of the rater/assessor. In chapters 6 and 7 we explored the cognitive processes that underlie judgment and decision making by raters when observing and assessing performance in the clinical workplace. Through the use of verbal protocol analysis we investigated how physician-raters (general practitioners) select and use observational data to arrive at judgments and decisions about trainees’ performance in a patient encounter.

Our findings with respect to information processing by raters during observation and evaluation of performance in single patient encounters are in line with many other studies in expertise research (Chi, 2006), indicating that levels of rater expertise are of major concern in WBA. Differences in raters’ knowledge structures and reasoning processes result from training as well as task-specific experience (i.e. rating experience). Compared to non-experienced raters, experienced raters seem to use more enriched processing, integrating observational data and case-specific, contextual information into comprehensive performance assessments. Increased activation and use of task-specific performance schemas in more experienced raters suggests that more experienced raters possess more and more sophisticated performance schemas compared to non-experienced raters. This may affect not only the nature of the feedback given to trainees, but also the accuracy of performance ratings (Lievens, 2001; Kerrins and Cushing, 2000).

Exploration of performance schemas used by raters in assessing trainee performance in patient encounters further supported our ideas of performance assessment being a specific application of ‘social perception’ for specific purposes, with raters behaving as ‘social perceivers’ (Klimoski and Donahue, 2001). When observing and evaluating trainee performance, raters interactively use (normative) performance theories (performance schemas), situation- or task-specific performance schemas and person schemas to arrive at judgments and decisions about performance effectiveness. An analysis of performance assessments by a large number of physician-raters enabled us to develop a collective, aggregate performance theory comprising 17 performance dimensions. Our findings, however, showed between-rater differences in the performance dimensions that were used in performance assessment, reflecting rater idiosyncrasy. This confirms findings in other domains regarding raters’ idiosyncratic (use of) performance schemata as a result of personal experiences, beliefs and attitudes (e.g. Uggerslev and Sulsky, 2008). Furthermore, our findings with respect to the building of person schemas are in line with person perception research, which shows that perceivers’ idiosyncratic interpretive processes may produce sharp differences in person perception (Mohr and Kenny, 2006).

In the context of WBA, differences in raters’ cognitive schemata may very well result in different representations of reality and thus different assessment outcomes. Major implications of our findings relate to the way we select raters for performance assessment, rater training and design of assessment systems. Overall, our studies provide support for the social-psychological, constructivist approach towards assessment as proposed in chapter 5. In order to improve the interpretation and use of work-based performance assessments, we therefore need to add to our traditional psychometric assessment frameworks and take alternative approaches which focus not so much on assessment outcomes as on the complex and interrelated cognitive, social,
emotional and contextual factors underlying judgment and decision making in assessment situations (Ferris et al., 2008; Levy and Williams, 2004).

Emergent themes

Performance assessment in professional competence development

Nowadays approaches towards assessment of (professional) competence are characterized by a shift away from individual assessment tools towards ‘purposeful arrangements of methods’ (programmes), on the basis of which valid inferences about professional competence can be made (Dijkstra et al., 2010; Van der Vleuten and Schuwirth, 2005; Hager et al., 1994). One of the challenges in designing assessment programmes, then, is how to combine assessment instruments and methods, reconciling the strengths and weaknesses of individual tools. Findings from our studies may provide useful insights for optimizing the use of performance assessments in assessment programmes to support development of professional competence.

Firstly, our studies clearly show the power of formalized assessments of performance -whether it be standardized performance-based tests or assessment of performance in work settings- for the guidance of competence development. The use of authentic, real-world assessments requires students to integrate relevant knowledge, skills and professional judgment performance, which guides learning towards expert forms of thinking that help students to integrate and link knowledge and skills to a broad range of clinical, social, cultural and other contextual factors (Hodges, 2006). Effectiveness and efficiency of competency-based education will therefore be enhanced by early introduction of performance assessments alongside early immersion in professional practice (Cooke et al., 2010). Although standardized assessment scenarios can be used in the early stages of professional education, ‘real world’ experiences and professional tasks should be used to guide and assess competence development whenever possible. Since standardized performance assessments are time consuming and expensive they should be used selectively for more advanced learners (e.g. for assessment of skills that cannot be assessed effectively or safely by other means).

Assessment programmes which aim to support life-long development of professional competence may take the shape of a ‘Z’, in line with current trends in (medical) curricula, emphasizing initial standardization and learning of fundamental aspects of performance, before gradually shifting towards practice-based assessment of ‘habitual’ performance in the real world. The shift in focus from traditional ‘school-based’ performance assessments towards high quality in-training WBA may thus facilitate a change of culture in which students and residents develop the habits of inquiry and improvement that promote excellence throughout a lifetime of practice (Cooke et al., 2010).

A major finding from our studies both on performance-based assessments and WBA during medical training is that combining formative and summative functions in assessment programmes seems to have a significant and positive impact on learning and competence development. This contradicts widely held views on assessment and common assessment
practices in which formative and summative assessments are strictly separated. Clearly, it is the summative assessment that drives student learning (chapters 2, 4). Effectiveness of formative assessment strategies in terms of influencing student learning is enhanced through well-considered incorporation of formative assessments in summative decision making strategies. However, in order to ensure that assessments effectively influence learning behaviours, trainees and assessees must have adequate opportunities to reflect on and actually use feedback for performance improvement. Optimizing the use of performance assessments for the development of professional competence therefore requires integration of these assessment strategies in longitudinal assessment programmes, in which frequent formative assessments provide opportunities for focused feedback and performance improvement as well as support summative decision making about competence achievement. This may have major implications for the way in which performance-based assessment strategies are used in medical training, shifting from typical ‘end-of-training’ assessment of performance (summative assessment of competence achievement) towards purposeful use of assessment for learning, embedded in teaching and learning. Clearly, this also implies the need for meaningful (narrative) feedback on performance as well as tools and procedures to facilitate documentation of performance, meaningful aggregation of performance data and fair and defensible decision making.

Our studies, however, also clearly demonstrate that a safe learning and working climate is a prerequisite for successful integration of formative and summative assessment functions. Both learners and assessors need to perceive the assessment system as ‘fit-for-purpose’, supporting honest feedback, accurate performance ratings and fair decision making. Essential preconditions are shared understanding of performance standards and criteria; assessment which is embedded in daily routines –fostering feedback seeking and receiving; both rater and ratee accountability as well as support in terms of time and training. Moreover, frequent and careful documentation of feedback and performance evaluations is essential, not only to support fair and credible decision making but also – and more importantly- to enhance the use of feedback for performance improvement. Moreover, effective use of feedback and performance assessments requires continuity in coaching and supervision, facilitating guided self-assessment and meaningful supervisor-trainee interaction relating to progress towards end-of-training learning goals. Major implications for in-training assessment, therefore, concern team collaboration in guidance and assessment of trainees’ competence development. Creation of small ‘communities of learning and practice’, in which both learners and teachers/supervisors have well-defined responsibilities and tasks, feel safe and have opportunities for prolonged engagement might be a feasible approach to realize effective learning, teaching and assessment in large-scale health care organizations.

Words versus numbers
Findings from our studies underline the importance of narrative comments in documenting performance assessments. Narrative feedback is central to learning, through identifying strengths and progress in learning as well as learning needs. Furthermore, our studies clearly indicate that narrative comments about performance are central to fair, i.e. credible and defensible decision making about competence achievement, through providing motivations and
arguments underpinning (numerical) performance scores – capturing the context in which task performance is embedded.

Moreover, research in industrial and organizational psychology suggests that a shift from numbers to words in performance assessment may have beneficial effects for the assessment process: the need to write down narrative comments which are used for summative decision making may increase involvement and commitment of assessors and may have implications for other feedback-related behaviours. Writing meaningful feedback and suggestions for improvement enables or forces assessors to structure their thoughts, to explicate (implicit) performance expectations and to think about how to formulate feedback in face-to-face situations. Generation of narrative comments might therefore be related to an enhanced feedback culture (Brutus, 2010). Field studies on the use of diaries in performance appraisal processes in industrial organizations furthermore showed that regular documentation of performance evaluations and feedback resulted not only in better feedback and more positive reactions to the appraisal process, but also in more accurate performance ratings (DeNisi and Peters, 1996; DeNisi et al., 1989).

We therefore argue that narrative feedback must play a significant role in assessment practices. It is clear that an increasing focus on producing narrative comments has implications for stakeholders in the assessment process (Brutus, 2010; Smither and Walker, 2004).

Clearly, a focus on narrative comments has implications for ratees / feedback recipients. In general, ratees tend to pay more attention to narrative comments, illustrating their increased potential for performance improvement – compared to quantitative feedback (Overeem et al., 2010; Smither and Walker, 2004). Reading comments about oneself, however, differs from interpreting quantitative results. Narrative feedback may elicit especially strong reactions, since they may convey a more personal focus: narrative comments are about and for the individual feedback recipient. As a consequence, feedback recipients may focus too heavily on comments that do not adequately represent their performance evaluation as a whole. Furthermore, interpretation of narrative feedback may be difficult. Comments may be vague, resulting in misunderstanding of feedback messages. Feedback which is received over time is likely to be composed of multiple comments of varying characteristics (e.g. positive and negative valence, referring to different competency domains, and containing various suggestions for development). This will considerably raise the complexity of the feedback message compared to single ratings. Complexity is exacerbated in the case of multiple evaluators for which some degree of divergence is often present (chapter 4, 7; Lance et al., 2008). From the perspective of the recipient, training and adequate coaching in interpretation and use of feedback is therefore needed to improve effectiveness of narrative comments for learning and competence development (Sargeant et al., 2008; Overeem et al., 2010; Luthans and Peterson, 2003).

Obviously, the production of meaningful and effective narrative feedback places high demands on raters’ feedback skills. Production of narrative comments requires more cognitive effort, challenging raters to put their thoughts into words and provide performance information without the directive prompts of items on a rating scale. Comments have to be useful and meaningful to have a positive effect on learning and performance. The relation between
characteristics of feedback and feedback effectiveness is complicated, though (Kluger and DeNisi, 1996; Hattie and Timperley, 2007; Shute, 2008). A study on the use of feedback following a multisource feedback programme indicated that characteristics of narrative comments (i.e. number, valence and focus (task/behaviour-focused versus trait-focused) of comments) determined actual changes in performance in ways which are similar to and as complex as face-to-face feedback (Smither and Walker, 2004). More importantly, recent research by Canavan et al. illustrated that the majority of comments in a multisource feedback programme for physicians-in-training lacked effective feedback characteristics (Canavan et al., 2010). Comments showed a low level of specificity and a significant amount of self-directed (trait-oriented) feedback. Research findings also suggest that there may be discrepancies between written feedback and face-to-face feedback communicated to the trainee (Murphy & Cleveland, 1995; Brutus, 2010). Especially when feedback is used for administrative purposes, raters are reluctant to write down negative feedback, whereas feedback which focuses on performance improvement may neglect positive feedback (Murphy & Cleveland, 1995; chapter 4). High-quality written feedback therefore calls for rater training which focuses not only on face-to-face feedback but also on how to write narrative comments. Accurate documentation of feedback has to be supported by well-designed and user-friendly tools and instruments that elicit narrative comments as well as procedures that promote use of feedback for performance improvement and trustworthiness in decision making (Van der Vleuten et al., 2010).

**Expert judgment: raters and ratees**

In more traditional assessments ‘assessment quality’ can be built in through quality assurance procedures in test construction, administration and psychometric analysis (Van der Vleuten et al., 2010; Norcini and McKinley, 2007). In performance assessments, and especially in WBA, quality of assessment is largely determined by the way assessment and assessment outcomes are used by those involved in the assessment system.

Performance assessments inherently rely on expert judgments by raters (assessors) who—ideally—have two types of expertise: considerable expertise in their professional domain and educational (assessor) expertise (Jones, 1999). Findings from studies as presented in this thesis show that assessor expertise affects information processing by raters when judging performance, in line with findings in industrial and organizational psychology (e.g. Lievens, 2001; Cardy et al., 1987). Different levels of rater experience may impact on feedback given to trainees and on accuracy of performance ratings (Lievens, 2001; Kerrins and Cushing, 2000). Indeed, it seems to be “the exercise of professional judgment that precludes competency-based assessment to become a mere ticking of checklists” (Jones, 1999; Brooks, 2009). Holistic judgments of performance, provided by professional judges, outperform detailed checklist ratings with respect to reliability and validity, thus improving utility of assessment outcomes (chapter 3; Cunnington et al., 1997; Regehr et al., 1998; Hodges et al., 1999).

In work-based performance assessments, raters are engaged in complex and unpredictable tasks in a context of time pressure and ill-defined or competing goals. Rater behaviour is furthermore shaped by their relationship with the ratee, interactions with others and by norms, values and other factors in the assessment context. Raters are continuously challenged to sample performance data; interpret findings; identify and define assessment criteria; and translate...
private judgments into sound (acceptable) decisions. Assessors in WBA contexts are therefore to be seen as ‘social perceivers’, who provide ‘motivated social judgments’ when evaluating performance (Murphy & Cleveland, 1995; Klimosky and Donahue, 2001; Levy and Williams, 2004). Depending on their interpretation of the assessment situation, assessors may decide what they perceive as ‘best’ rather than what is ‘right’ in an absolute sense. The concept of professional judgment in assessment situations is therefore very similar to the concept of professional judgment in other domains such as the health professions (Coles, 2002). Obviously, development of professional judgment requires education, i.e. rater training focusing on knowledge and skills. It first and foremost, however, develops through experience, deliberation and deliberate practice. This inevitably implies that expert judges develop individual performance constructs and ‘practical wisdom’ underlying assessments of performance. Relying on expert judgment, therefore, implies relying on “idiosyncratically constructed realities unique to individual judges” (Van der Vleuten et al., 2010).

Our approach towards professional judgment as described above may have implications for the way we select and train our judges. As any professional expert, (professional) judges need “to engage in the appreciation of their practice “, which entails critical reflection on not just what they do, but also on ‘what they are and the attitudes, beliefs that underpin their actions” (Coles, 2002). Brief, one-off training sessions will not do if we aim for excellence in performance assessments; development of professional judgment requires long-term support, coaching and feedback as well as reflection on strategies used in judging (complex) performance in different (ill-defined) contexts (Ericsson, 2004; Coles, 2002).

Key to the effectiveness of performance assessments is rateee engagement, as this is crucial for acceptance and use of feedback. Research evidence indicates that assessment systems which are designed according to principles of ‘due process’, paying explicit and specific attention to rateee involvement in the assessment process, enhance satisfaction with assessment outcomes. In health professions education as well as in other domains, self-ratings or self-assessments of performance are considered to enhance reflection on performance and the use of feedback for performance improvement. In other words, they are vital to self-directed learning and professional self-regulation (Duffy and Holmboe, 2006). There is conclusive evidence, however, that people are bad self-assessors (Davis et al., 2006; Gordon, 1991; Kruger & Dunning, 1999), casting doubt on the ability of professionals to function as self-regulating professionals (Eva and Regehr, 2005). Recent research indicates that under conditions of coaching, mentoring and a safe learning environment, comparison of self-ratings with expert scores and external feedback may stimulate reflection on performance, as well as enhance acceptance and use of external feedback for high quality learning and performance improvement (Overeem et al., 2010; Luthans and Petersen, 2003). Nowadays approaches towards self-assessment therefore adopt models of ‘directed’ self-assessment or ‘guided’ self-assessment in which trainees are encouraged to compare their self-perceptions of performance with external feedback, and reflection on feedback and use of feedback for practice improvement are facilitated (Sargeant et al. 2008; Davis, 2009; Duffy and Holmboe, 2006). Introducing (‘guided’) self-assessments at early stages of medical training, incorporating them into performance assessment strategies, may thus support
the development of students and trainees into professionals who are able to make ‘informed judgments’ about their own performance to improve professional practice.

**Assessment in context**

Efforts to understand performance assessments typically tend to take a psychometric approach, focusing on assessment design, instrumentation and rating scales, or on raters and ratees as individuals in the assessment process (Norcini and McKinley, 2007). Without any doubt, past research has contributed significantly to improvement of assessment in health professions education. Based on research presented in this thesis as well as a wealth of research findings in various domains (Murphy and Cleveland, 1995; DeNisi, 1996; Hawe, 2003; Levy and Williams, 2004) we argue, however, that performance assessment can be completely understood only *in situ*. Both rater and ratee behaviours are framed within the context in which assessment processes take place. Social, political, cognitive, emotional, and relationship factors in the (educational or work) context determine assessment outcomes, and it is important to carefully consider these factors when investigating and interpreting performance assessments (Levy and Williams, 2004; Ferris et al., 2008). A simplified model of the performance assessment process is presented in Figure 8.2. Central to the model are cognitive processes in which raters are engaged when making judgments and decisions about ratee performance. Judgment and
decision making are influenced by many factors, and actual ratee performance is just one out of many. Features of the assessment system, organizational norms, values as well as rater characteristics (e.g. experience, mood, personality) influence processes underlying judgment and decision making, thus affecting assessment outcomes.

It is furthermore important to realize that raters and ratees are, to some extent, co-creators of the assessment context: they are active participants in, interact with and shape the assessment context through their behaviours within this context. For instance, a study by Tziner et al. (2001) showed that raters who believe or experience others to inflate and distort their ratings are likely to themselves inflate and distort – thus influencing the organization’s feedback culture, norms and values.

Key issues in meaningful work-based performance assessments seem to be related to accountability, trust and acceptance of assessment systems as being fair and ethical. Research findings show that accountability pressures on raters as well as ratees may increase acceptance, commitment, accuracy in performance ratings and performance notes as well as use of feedback for performance improvement (Mero and Motowidlo, 1995; Mero et al., 2003; Walker and Smither, 1999). Elements of due process (adequate notice, fair hearing and judgment based on evidence) may further enhance perceptions of fairness in work-based performance assessments (Taylor et al., 1995; Erdogan et al., 2001). Our research findings confirm the importance of: clear communication and discussion of performance theory (criteria, standards) and regular, timely feedback (adequate notice); transparency in decision making procedures, rater credibility and ratee involvement in the assessment process (fair hearing) as well as regular documentation of performance evaluations and notes explaining performance ratings and providing opportunities for discussion and reflection (evidence-based judgments). Principles of due process and accountability seem to be very much in line with guidelines for ‘directed self-assessment’ (Sargeant et al., 2008). They therefore seem to be key in fostering a ‘feedback culture’ in which individuals at all levels of competence development are encouraged and facilitated to seek external feedback, to assess and reflect on their own performance and use feedback for performance improvement.

Implications for further research

The research presented in this thesis was aimed at furthering our understanding of performance assessment in health professions education. Using conceptual frameworks from social cognition, expertise research and judgment and decision making, we propose approaches to assessment from a constructivist social-psychological perspective. The results presented in this thesis call for further research in various areas.

Firstly, the model as presented in this chapter needs to be confirmed and refined through additional research in different performance assessment contexts. Our assessment model indicates that, in order to better understand outcomes of work-based performance assessments, we need to focus on the interrelationship between rater, ratee, and the assessment context. It is
clear that more empirical work should be conducted to isolate and investigate (understand) these relationships in the setting of health professions education—at all levels of the medical education continuum. Extrapolating from research findings in industrial and organizational psychology, one might assume that a large set of variables in the social context of work-based assessment may affect rater and ratee behaviours and thus assessment outcomes (Levy and Williams, 2004; Ferris et al., 2008). Although the list of potential research questions is virtually endless, highly relevant research domains concern acceptance of and trust in performance assessment systems. This implies a shift in focus from psychometric analysis of assessment outcomes to investigation of both ratees’ AND raters’ perceptions of and reactions to assessment systems. Performance assessment in the real world is arguably an emotional event, both for ratees and raters (Ferris et al., 2008; Brutus, 2010). Emotional reactions to performance assessment are likely to be shaped by individual differences (personality characteristics, emotional intelligence, political or leadership skills) as well as the quality of rater-ratee relationships. In order to better understand the effectiveness or ineffectiveness of performance assessments, future research should therefore focus on the affective outcomes of performance assessment systems, and how affective reactions may impact not only on the performance of the individual ratee but also on rater-ratee work relationships and performance as a team.

Clearly, we also need further research to identify how contextual factors impact on beliefs and attitudes toward assessment—thus influencing raters’ and ratees’ behaviours within assessment systems.

For instance, performance assessment systems can be seen as accountability mechanisms within educational or organizational (work) contexts. Although accountability has been shown to have positive effects on quality of performance evaluations and use of feedback for performance improvement, accountability may also raise ‘tension’ and resistance. Dysfunctional behaviour (e.g. distortion of performance ratings, strategic selection of raters and moments of observation, raters and ratees ‘playing the game’) may be the result, reflecting compliance rather than internalization of principles underlying the performance assessment system. Further research addressing the balance between accountability and freedom or self-control within assessment systems may support design and implementation of effective assessment systems.

Central to the implementation of performance assessment systems, especially in work settings, is improvement of performance and—ultimately—improvement of health care quality. A leading research question, therefore, concerns the impact of performance assessment systems on the quality of care. Providing direct empirical evidence to support this relationship, however, will be very difficult—if not impossible. However, starting from the notion that feedback (seeking) and deliberate practice are key to performance improvement, an important area of research might be the impact of performance appraisal systems on the feedback culture within organizations.

Some of our findings may redirect research efforts towards instrumentation, assessment process and rater training.

For instance, an approach towards assessment which heavily relies on narratives is bound to have implications for collection and communication of performance information. Raters will be challenged to provide narrative comments that are motivating and meaningful in guiding
competence development as well as useful in decision making about competence achievement. Consequently, research questions may concern the design of assessment instruments: do we need quantitative ratings in combination with narrative feedback; and if so: when and how? How can we facilitate raters to provide comments that are meaningful and useful, supporting learning as well as credible decision making? How does this impact on quality of performance assessments, for instance in terms of decreased leniency (Brutus, 2010)? Furthermore, we need research exploring meaningful aggregation and interpretation of qualitative performance data for decision making, across assessment sources and occasions. Conceptual frameworks derived from qualitative research as well as judgment and decision making theories have been proposed (Van der Vleuten et al., 2010), but more research is needed to fully appreciate and optimize the utility of assessment outcomes in work-based settings. Finally, our findings indicate that information processing in judgment and decision making may differ, depending on level of rater expertise. Raters’ performance constructs develop over time, influenced by rater training, socialization and personal experience in assessment of performance. In order to better understand processes underlying judgment and decision making by raters with different levels of rater experience, future research efforts should include exploration of raters’ performance theories and factors influencing raters’ expertise development. Additional research is needed to investigate the relationship between rater expertise and feedback given to ratees and effects on performance improvement.

In conclusion

This thesis aimed to contribute to a better understanding of performance assessment. Our studies partly reproduced and confirmed research findings on performance assessments in medical education. Research as presented in this thesis contributes to our current understanding of performance assessment through focusing on assessment as embedded in the larger context of the educational and work environment. Assessment outcomes are determined by complex and interrelated processes, influenced by cognitive, emotional, political and interrelationship factors. In order to better understand judgment and decision making in performance assessment systems, we therefore need to take approaches which shift from an exclusive psychometric focus on assessment outcomes towards social-psychological, constructivist assessment approaches. Within the complex social context of performance assessments, both raters and ratees have to develop into expert judges, who are able and willing to adequately use professional judgments in assessment processes, fostering and judging professional competence development.
References


SAMENVATTING

Summary in Dutch
Naar de top van de piramide: het toetsen van ‘performance’

Toetsprogramma’s in het medisch onderwijs worden gekenmerkt door een toenemende aandacht voor het evalueren van ‘performance’. ‘Performance’ wordt daarbij gedefinieerd als het handelen in de realistische beroepspraktijk, of in situaties die een min of meer levensechte afspiegeling vormen van die praktijk. Vanwege het ontbreken van een goed Nederlands equivalent zal in deze samenvatting de Engelse term ‘performance’ gebruikt worden, en als afgeleide daarvan ‘performance-based assessment’ (doorgaans gebruikt voor toetsen op het niveau van Miller’s ‘shows-how’) en ‘performance assessment’ (‘does’-niveau in Miller’s piramide). Toetsing in medisch onderwijs richt zich steeds meer op de bovenste lagen van de piramide van Miller.

De belangrijkste achterliggende redenen voor veranderingen in de toetsing zijn de invoering van competentiegericht onderwijs; veranderende ideeën over wat het betekent om een ‘goede dokter’ (of breder: zorgverlener) te zijn en te blijven; en verantwoording die over kwaliteit van zorg en kwaliteit van opleiden moet worden afgelegd. Het volstaat niet langer om oordelen over professionele competentie uitsluitend te baseren op traditionele kennisgerichte toetsen of op impliciete, globale impressies van het functioneren in de praktijk. Daarnaast is er overtuigend bewijs dat regelmatige evaluatie van en feedback op performance essentiële voorwaarden zijn voor de ontwikkeling van professionele competentie en behoud van expertise. Expliciete, geformaliseerde toetsing van performance zal daarom een niet meer weg te denken plaats gaan innemen in medische opleidingsprogramma’s, maar ook in de carrière van professionele zorgverleners.

![Diagram van Miller’s piramide](Figuur 1 De piramide van Miller)
Dit proefschrift bestaat uit een aantal studies naar aspecten van toetsing van performance. Hoofdstukken 2 en 3 richten zich op gestandaardiseerde ‘performance-based assessments’, hoofdstukken 4 tot en met 7 beschrijven studies naar het hoogste niveau in de piramide. Centrale onderzoeksvragen betroffen:
- Bruikbaarheid van ‘performance-based assessments’ in relatie tot grootte van het taakdomein en mogelijkheden tot verbetering van de bruikbaarheid van deze toetsen;
- Bruikbaarheid van een semigestandaardiseerde vorm van ‘performance assessment’ in de klinische context;
- Processen die ten grondslag liggen aan oordeels- en besluitvorming bij ‘performance assessment’, en meer er specifiek:
  - Factoren die van invloed zijn op oordeels- en besluitvorming in realistische leer-werksituaties;
  - Cognitieve processen die ten grondslag liggen aan expertoordelen (d.w.z. oordelen gegeven door professionals uit het werkveld) bij het beoordelen van performance in de klinische context.

**Hoofdstuk 2 en hoofdstuk 3** beschrijven de ontwikkeling, implementatie en evaluatie van een instrument voor gestandaardiseerde beoordeling van performance, als maat voor professionele competentie, in de Opleiding Verloskunde aan de Academie Verloskunde Maastricht (voorheen: Vroedvrouwenschool Kerkrade / Maastricht). Gebaseerd op ervaringen en onderzoeksbevindingen uit het geneeskundeonderwijs werd een stationstoets ontwikkeld, de Verloskundige CompetentieToets (VCT), bestaande uit 6 stations van elk 30 minuten. Met behulp van een vragenlijst, afgenomen bij studenten en examinatoren, werden acceptabiliteit en onderwijsconsequenties van de toets in kaart gebracht. De betrouwbaarheid van de toetscores werd onderzocht met behulp van generaliseerbaarheidstudies. Tevens is nagegaan hoe de toetsbetrouwbaarheid - en daarmee toetsesfectiviteit - beïnvloed wordt door a) het gebruik van globale beoordelingsschalen in plaats van traditionele taakspecificie criterialijsten; en b) de keuze voor een beheersingsgericht perspectief bij score-interpretatie (betrouwbaarheid van zak-slaagbeslissingen).

Zoals te verwachten, is de acceptabiliteit van de VCT hoog. Zowel studenten als examinatoren (stafleden en praktiserend verloskundigen) zijn van mening dat de VCT de kwaliteit van het toetsprogramma aan de opleiding verhoogt, door te focussen op de integratie van kennis, vaardigheden en inzicht in verloskundig handelen. De toets heeft een positieve invloed op het leergedrag van studenten, door het stimuleren van een casus- of probleemgestuurde (in plaats van leerboekgestuurde) studieaanpak. Daarnaast heeft de ontwikkeling en invoering van de toets een positief effect op het opleidingscurriculum, door het stimuleren van discussies over de essentie van professionele competentie, en de consequenties voor inhoud en opzet van het onderwijs. De resultaten tonen aan dat ook in het relatief smalle verloskundige taakdomein (vergeleken met het bredere geneeskundige domein) betrouwbaarheid van scores bij stationsexamens problematisch is, waarbij inhoudsspecificiteit als belangrijkste verklarende factor blijft bestaan. De geschatte toetsbetrouwbaarheid op basis van de criterialijstscores is matig (generaliseerbaarheidscoëfficiënt = 0.48); een toetsduur van meer dan 10 uur is nodig voor het bereiken van acceptabele betrouwbaarheid (.80). Toepassing van globale oordelen leidt
Samenvatting
tot een toename van toetsbetrouwbaarheid (generaliseerbaarheidscoëfficiënt = 0.61), en deze
can nog verder verhoogd worden door te kiezen voor een beheersingsgericht perspectief bij
score-interpretatie. De betrouwbaarheid van zak-slaagbeslissingen op basis van criterialijstscores
is 0.85; op basis van globale oordelen bedraagt deze 0.95. Deze bevindingen laten zien dat
weloverwogen inzet van ‘experts’ (professionals) en expertoordelen (‘professional judgement’)
bij beoordelen van performance kan bijdragen aan kwaliteit en bruikbaarheid van
toetsresultaten.

Hoofdstuk 4 presenteert een focusgroeponderzoek naar ervaringen, opvattingen en percepties
van studenten met betrekking tot systematische en semigestandaardiseerde toetsing van
performance gedurende stages in de Opleiding tot Verloskundige. Centraal hierin staat het
gebruik van zogenoemde ‘observatieboeken’ (observational diaries), waarin gedetailleerde
checklijsten zijn opgenomen voor frequentie (dagelijkse) schriftelijke verslaglegging van feedback
en evaluatie van performance. De observatieboeken beogen daarmee zowel het leren van
studenten te sturen (formatieve functie van toetsing) als wel beslissingen over het bereikte
competentieniveau te onderbouwen (summatieve functie). Deze integratie van toetsfuncties lijkt
in strijd met klassieke opvattingen waarin gepleit wordt voor een strikte scheiding, onder andere
vanwege ongewenste effecten van combinatie van formatieve en summatieve toetsfuncties op
student- en docentgedrag. In drie focusgroepen, afkomstig uit twee opeenvolgende cohorten
vierdejaarsstudenten, is uitvoerig gediscussieerd over effecten van deze vorm van stagetoetsing
en observatieboeken op student- en docentgedrag, en de bruikbaarheid van informatie ten
behoeve van summatieve beslissingen.
Resultaten suggereren dat geformaliseerde toetsing op de werkplek, tijdens stages, een positief
effect heeft op studentleergedrag. Frequentie documentatie van feedback en evaluatie van
performance aan de hand van helder gedefinieerde criteria en standaarden stuurde de
ontwikkeling van professionele competentie in de gewenste richting, en verhoogt gevoelens van
self-efficacy gedurende de stage. Systematische en gestructureerde evaluatie van performance,
het vastleggen van feedback speelt daarmee een belangrijke rol bij het expliciteren van
leerprocessen op de werkplek. Het positief sturend effect op professionele ontwikkeling kan nog
worden vergroot door studenten een actieve rol toe te bedelen in het proces van toetsing,
bijvoorbeeld in de vorm van zelfevaluaties. Gedetailleerde criterialijsten, hoewel zeer bruikbaar
bij aanvang van stages, hebben naarmate de expertise van studenten toeneemt een remmend
effect op het leren. Leren wordt vooral gestuurd door narratieve feedback en toetsinstrumenten
die het mogelijk maken aan te sluiten bij de persoonlijke ontwikkeling van de student. Naast het
inzichtelijk maken van sterke kanten en verbeterpunten in het functioneren, is narratieve
feedback ook onmisbaar in zorgvuldige besluitvorming: narratieve feedback geeft betekenis aan
nummerieke scores door het expliciteren van de overwegingen en argumenten van de supervisor-
beoordelaar bij de totstandkoming van zijn\(^4\) oordeel over kwaliteit van handelen.
Resultaten laten zien dat formatieve en summatieve functie van toetsen elkaar versterken.
Enerzijds stimuleert het gebruik van performance-evaluaties voor summatieve doeleinden
(onderbouwing van beslissingen) de formatieve functie, door directe observatie; het geven van
\(^4\) Daar waar ‘hij’ of ‘zijn’ staat kan ook ‘zij’ of ‘haar’ gelezen worden.
Samenvatting

betekenisvolle feedback en door documentatie van deze feedback te bevorderen. Anderzijds zijn frequente formatieve evaluaties en de schriftelijke verslaglegging daarvan, noodzakelijke voorwaarden voor verdedigbare besluitvorming. Op basis hiervan concluderen wij dat formatieve en summatieve functies in longitudinale toetsprogramma’s, bijvoorbeeld voor het beoordelen van functioneren op de werkwkplek, gecombineerd kunnen en moeten worden. Deze bevindingen ondersteunen recente ontwikkelingen in de medisch-specialistische vervolgonderwijs, waarin besluitvorming over voortgang in de opleiding onderbouwd wordt met informatie uit formatieve evaluaties van het functioneren van de AIOS, verzameld in het AIOS-portfolio.

Een veilig leerklimaat vormt de basis voor succesvolle implementatie van deze vorm van toetsing op de werkwkplek. Inbedding van toetsing in de dagelijkse werkroutines; heldere criteria en standaarden; eerlijke en betekenisvolle feedback; en zorgvuldige besluitvormingsprocedures waarin de lerende een actieve rol vervult, zijn de sleutelfactoren die bepalend zijn voor het vertrouwen in en commitment met het toetsprogramma – en daarmee voor adequaat gebruik van toetsresultaten voor verbetering van performance.

Hoofdstuk 5 geeft een overzicht van factoren die ten grondslag liggen aan de totstandkoming van expertoordelen bij het beoordelen van performance op de werkwkplek. Expertoordelen (‘professional judgements’) zijn inherent aan beoordelen van performance op de werkwkplek. Onderzoek, met name vanuit psychometrisch perspectief, toont aan dat betrouwbaarheid en validiteit van werkwkplekbeoordelingen regelmatig te wensen overlaat. Een aantal van de waargenomen problemen, zoals halo-effecten, gebrekkige differentiatie tussen studenten en cijferinflatie, wordt toegeschreven aan meetfouten als gevolg van ontoereikende bekwaamheid van beoordelaars. Pogingen om de betrouwbaarheid en validiteit van werkwkplekbeoordelingen te verbeteren zijn dan ook vaak gericht op vaardigheidstraining van beoordelaars, teneinde de kwaliteit van beoordelaars als meetinstrument te optimaliseren. Deze trainingen blijken echter een beperkt effect te hebben. Een voor de hand liggende vraag is dan ook welke (andere) factoren ten grondslag liggen aan gedrag van beoordelaars bij werkwkplekbeoordelingen. Op basis van een literatuurstudie en onderzoeksresultaten uit verschillende disciplines worden in hoofdstuk 5 een aantal factoren geïdentificeerd die van invloed (kunnen) zijn op gedrag van beoordelaars. Een drietal categorieën van factoren wordt besproken:

1. cognitieve factoren;
2. opvattingen over performance en kwaliteitsstandaarden;
3. motivationele factoren.

Uit onderzoek blijkt dat beoordelaars verschillende cognitieve strategieën kunnen hanteren bij hun oordeelsvorming, en dat de kwaliteit en bruikbaarheid van de beoordeling afhankelijk is van de gevolgde strategie. Beoogde functies van toetsing en eigenschappen van toetsinstrumenten beïnvloeden cognitieve processen bij beoordelaars. Beoordelaars hebben daarnaast vaak uiteenlopende opvattingen over wat doorslaggevend is voor kwaliteit van performance en welke eisen gesteld moeten worden aan lerenden. Beargumenteer is dat expertoordelen per definitie idiosyncratisch zijn: beoordelaars vormen hun eigen normatieve ‘performance theorie’ op basis van training, socialisatie en werkervaring. Veel studies tonen daarnaast aan dat besluitvorming (d.w.z. de uiteindelijke, gecommuniceerde beoordeling) vooral wordt beïnvloed
door factoren in de werkomgeving. Waarden en normen m.b.t. toetsing, persoonlijke doelstellingen bij het beoordelen van iemands performance, eventuele gevolgen van een negatieve beoordeling voor de beoordelaar, rolconflicten en af te leggen verantwoording (accountability) hebben alle een belangrijk sturend effect op beoordelaarsgedrag en daarmee op kwaliteit van oordelen. Opvattingen, motieven en intenties van beoordelaars lijken daarmee belangrijker verklaringen voor de tekortkomingen in werkplekbeoordelingen dan de vaak veronderstelde onbekwaamheid. Beoordelaars zijn geen passieve meetinstrumenten en moeten ook niet als zodanig behandeld worden. Integendeel, beoordelaars zijn actieve verwerkers van informatie. Beoordelingen komen tot stand op basis van cognitieve processen die vergelijkbaar zijn met bijvoorbeeld klinisch redeneren: selectieve en doelgerichte informatieverwerking, interpretatie van informatie, integratie van informatie en doelgericht gebruik van de opgeslagen informatie bij het nemen van beslissingen. De uitkomsten van het besluitvormingsproces worden bepaald door percepties van de beoordelaar, interpretaties van de beoordelingstaak in de context van de eigen werkomgeving, en de consequenties van besluitvorming voor betrokken partijen. Aandacht voor de context waarin ‘performance assessments’ plaatsvinden en hoe deze van invloed is op beoordelaargedrag, is daarom cruciaal bij kwaliteitsverbetering van deze vorm van toetsing.

Voortbouwend op onze beschouwing van onderzoeksbevindingen uit diverse disciplines pleiten wij voor een constructivistische, sociaal-psychologische benadering van ‘performance assessments’ waarin inzichten uit onder andere sociale cognitie en besluitvormingstheorieën geïntegreerd worden, in aanvulling op de traditioneel psychometrische benadering van toetsing.

Hoofdstukken 6 en 7 presenteren onderzoek naar cognitieve processen bij beoordelaars die ten grondslag liggen aan oordeels- en besluitvorming bij het beoordelen van performance in de klinische setting. Door gebruik te maken van hardopdenkmethoden en analyse van verbale protocollen hebben we onderzocht hoe supervisor-beoordelaars (huisartsopleiders) informatie selecteren en gebruiken bij directe observatie en beoordelen van studentperformance in een authentieke taak (consultvoering). Gebruikmakend van theoretische inzichten uit onderzoek naar expertise- en expertise-ontwikkeling, hebben we tevens gekeken naar verschillen in cognitieve processen bij ervaren (N = 18) en onervaren (N = 16) beoordelaars. Resultaten komen grotendeels overeen met bevindingen uit expertiseonderzoek in andere domeinen. Vergeleken met onervaren beoordelaars, maken ervaren beoordelaars significant meer gebruik van ‘inferenties’ (interpretaties ofwel bewerkingen van informatie). Niet-ervaren beoordelaars daarentegen genereren in het hardopdenken meer letterlijke beschrijvingen van wat ze hebben waargenomen. Ook besteden ervaren beoordelaars in hun oordeelsvorming meer aandacht aan situatiespecifieke of taakgebonden factoren die bepalend zijn voor kwaliteit van handelen in betreffende situatie. Al in een vroeg stadium van het proces van oordeelsvorming wordt relevante taakspecifieke informatie geïntegreerd met performance informatie, om zo te komen tot een betekenisvolle interpretatie van en omvattend oordeel over kwaliteit van handelen. De toename in het gebruik van taakspecifieke performance-schema’s bij ervaren beoordelaars suggerereert dat ervaren beoordelaars beschikken over meer, of meer verfijnde performance-schema’s dan onervaren beoordelaars. De gevonden verschillen in informatieverwerking hebben
mogelijk consequenties voor de kwaliteit (accuratesse) van de beoordeling, maar ook voor de kwaliteit van de feedback aan de lerende.


Hoofdstuk 8 geeft een samenvatting van de voorgaande hoofdstukken in het proefschrift. Vervolgens wordt ingegaan op een viertal thema’s die min of meer nadrukkelijk uit de studies naar voren komen.

Samenvatting

verantwoordelijkheden voor alle betrokkenen, biedt mogelijkheden om het leren, begeleiden en beoordelen op de werkplek te optimaliseren.

Een tweede thema dat uit onze studies naar voren komt, is de noodzaak tot meer nadruk op narratieve feedback. Narratieve feedback is niet alleen essentieel bij het sturen van het leerproces, maar ligt ook ten grondslag aan zorgvuldige besluitvorming –als motivering bij numerieke scores, inzicht gevend in persoonlijke performance theorieën van de beoordelaar en zijn interpretatie van de beoordelingstaak. Onderzoeksbevindingen, met name in de arbeids- en organisatiepsychologie, suggereren tevens dat het benadrukken van narratieve feedback een positief effect kan hebben op de feedbackcultuur in organisaties en acceptabiliteit van ‘performance assessments’. Toepassen van narratieve feedback in toetsing van performance stelt eisen aan alle betrokkenen. Adequate interpretatie en gebruik van de feedback vereist daarom training en coaching van zowel gevers als ontvangers van feedback. Documentatie van feedback dient daarnaast te worden gefaciliteerd door inzet van (gebruiksvriendelijke) toetsinstrumenten en procedures die waarborgen dat optimaal gebruik wordt gemaakt van narratieve feedback voor ontwikkeling en beoordeling van professionele competentie.


In thema 4 beargumenteren we dat gedrag van beoordelaars en beoordeelden, en daarmee uitkomsten van ‘performance assessments’ bepaald worden door de context waarin de
Samenvatting


Tenslotte presenteren we aan het eind van hoofdstuk 8 een aantal suggesties voor verder onderzoek. Centraal staat de vraag hoe een systeem voor toetsing van performance kan bijdragen aan continue verbetering van kwaliteit van handelen en -  uiteindelijk - kwaliteit van zorg. Uitgangspunt hierbij is de notie dat feedback en ‘deliberate practice’ doorslaggevend zijn in de ontwikkeling van expertise. Relevante onderzoekvragen betreffen de impact van ‘performance assessments’ op de feedbackcultuur in organisaties en omgekeerd, de invloed van contextfactoren op percepties, opvattingen en gedrag van actoren in deze toetssystemen.