Climbing the pyramid : towards understanding performance assessment

CLIMBING THE PYRAMID

Towards Understanding Performance Assessment

To my parents
To my godmother and grandma

The research presented in this thesis was carried out at

Academie verloskunde
Maastricht

and

Maastricht University Leading in Learning!

in the School of Health Professions Education

SHE

ISBN 978-90-816534-0-4

Layout by: Maartje Gijselaers
Cover design by: Bonné +Jan
Printed by: Drukkerij Pietermans

CLIMBING THE PYRAMID

Towards understanding performance assessment

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van de Rector Magnificus,
Prof. Mr. G.P.M.F. Mols,
volgens het besluit van het College van Decanen
in het openbaar te verdedigen
op donderdag 8 september 2011 om 16.00 uur

door

Maria Johanna Bernadette Govaerts

Promotores

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

Co-promotor

Dr. Ir. A.M.M. Muijtjens

Beoordelingscommissie

Prof. Dr. R.P. Koopmans (voorzitter)
Prof. Dr. Th.J. ten Cate (Universiteit Utrecht)
Prof. Dr. F. Scheele (Vrije Universiteit Amsterdam)
Prof. Dr. M.S.R. Segers
Dr. B.H. Verhoeven

Table of Contents

CHAPTER **1**

Introduction

Introduction

Medical education can be conceptualized as the complex of processes by which a student (trainee) changes from a layman to a competent professional in the medical domain, implying changes from unknowing to knowing, from unskilled to skilled, from incompetent to competent (Abrahamson, 2000). In undergraduate as well as postgraduate education, medical educators are continuously challenged to create and provide learning environments which effectively and efficiently support and guide trainees' competence development. At the same time, they have to develop strategies to ensure credible and defensible decision-making about achievement of competence, allowing trainees to proceed to the next level of education, responsibility and -eventually- independent practice. They have to ensure that newly qualified professionals have achieved standards of competence that are acceptable to the profession and the community (Dauphinee, 2002).

Obviously, high quality assessment is important throughout medical education. Assessment can be defined as "any (systematic) method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects or programs" (Standards for Educational and Psychological testing, 1999, p. 172). Assessment outcomes, reflecting trainee performance in relation to educational goals, not only provide information to substantiate decisions about competence achievement for purposes of selection or promotion, but can also give a focus to further learning (Crossley et al., 2002; Epstein, 2007). In fact, some argue that assessment is the main driving force behind student learning (Frederiksen, 1984; Struyven et al., 2003) and it is generally acknowledged that assessment of and feedback on performance are key to development (and maintenance) of expertise (Ericsson, 2004).

Although the formative function of assessment and its impact on the quality of student learning has long been acknowledged, issues of accountability, fairness and equivalence in the context of summative assessment have dominated assessment development and research in medical education until now: the public expects and demands that practitioners who provide health care are competent.

In assessment we trust

> *"Over 95% of students who enter medical school will graduate with their medical degree and go on to practice medicine. By definition, they have fulfilled their requisite competencies. Yet anyone who works in medical education knows that we wilfully graduate medical students every year whom we would never trust with our family members' medical care". (p. 100)*
>
> *"I, like all practicing physicians, know other physicians whom I wouldn't let near me or my family members, regardless of the number of pieces of paper they might sport." (p. 95)*
>
> *M. Brooks, 2009*

> *"Any medical student or physician –who has had professional contact with colleagues- knows which of the colleagues is a good physician, which is not equal to a 'competent' physician! ………. The best surgeons, for instance, are not necessarily those who have the best technical skills (know best how to do something), but rather those who know what to do, when, and why, and especially when <u>not</u> to do something". (pp. 97, 99)*
>
> *M. Brooks, 2009*

Although this might be easily dismissed as too cynical a judgment of assessment in the medical profession, there is some research evidence to support Brooks' remarks. Medical students, upon graduation, too often show (substantial) deficiencies in knowledge and clinical skills (Reilly, 2003; Mangione and Nieman, 1997; Wilkerson and Lee, 2003). Moreover, even if deficiencies in performance are detected, students do not always receive honest and meaningful feedback, nor are they encouraged to enrol in remediation programmes (Hauer et al., 2009; Ginsburg et al., 2000; Van der Hem-Stokroos et al., 2001). Not only does this undermine the public's (and the profession's) confidence that newly qualified doctors are actually competent, it also –perhaps even more importantly- deprives students of opportunities to improve performance. Findings are the more alarming since research has shown that performance during medical school is predictive of future practice performance (Papadakis et al., 2005; Tamblyn et al., 2002).
Similar problems have been described in postgraduate medical education, which in fact is where trainees nowadays are prepared for practice. Research by Williams and Dunnington, for instance, showed that most residents receive extremely high performance ratings (more than 85% of residents' performance is rated as 'excellent' or 'very good'), whereas a negligible number of residents are identified as having (serious) problems with respect to performance (Williams and Dunnington, 2004). Lack of direct observation and hence of meaningful feedback as well as limited documentation of performance evaluations, are considered to be major reasons underlying these problems (Williams et al., 2005).

Finally, research findings suggest that as many as 10% of physicians will demonstrate significant deficiencies in knowledge or skills at some point in their careers (Leape and Fromson, 2006; Choudry et al., 2005; Bolk, 2000). Thus, physician performance failures are not rare and pose substantial threats to patient safety and welfare (Leape and Fromson 2006). Lack of direct observation and feedback are well-known features of the clinical educational setting, but it is even more common for doctors to fail to provide frequent and honest performance feedback to their peers - whether it be praise or concerns about performance -, let alone to officially report incompetent physicians (Ginsburg et al., 2000; Donaldson, 1994). This is tragically illustrated by high profile cases of alleged medical incompetence and misconduct (e.g. Bristol case in the UK; Swango case in the U.S.A.; Radboud case in the Netherlands).

Educators have been struggling with assessment for decades. Although it is beyond any doubt that the majority of health care professionals and graduates in medical education are 'fit-for-practice', the issues described above indicate that our struggle with assessment is not over yet, and chances are that it will continue for the foreseeable future.

Several key issues emerge from the literature as cited above:

- How can we improve assessment of performance? Do our assessment methods enable us to make valid inferences about professional competence and (future) practice performance? What should be assessed, and how? How can we improve feedback processes, and the use of feedback for performance improvement?
- What are the main barriers to credible and defensible decision making on 'professional competence' in the educational setting, and in the setting of professional practice? Why don't we fail medical students and / or residents whose performance is substandard? How can we improve the use of assessment results for feedback and decision making?
- Are professionals competent to assess competence? What do we actually know about professional judgment in assessment situations? How do professionals arrive at judgments about performance and competence, and how do we explain discrepancies between private judgments and public decision making –as illustrated in the paper by Brooks?

The problems described above are not unique to medical education. 'Failure to fail' and reluctance of supervisor-professionals to act as gatekeepers to the profession have been described in other domains as well (Hawe, 2003; Ilott and Murphy, 1997), as has the need for adequate recertification procedures (Lysaght and Altschuld, 2000).

The present thesis aims to address some aspects of the complexity of assessment as implied in the above questions. The studies included in this thesis are all situated in the domain of health professions education, but findings are relevant to other professional domains as well.

In the following we first describe recent developments in assessment in medical education. Current insights into assessment of professional competence and performance, as well as the main barriers to effective implementation are identified. The studies addressing the research questions that emerged from our framework are summarized in a thesis outline.

Where are we: standardization in diversity

Until the 1950s-1970s assessment in medical education largely relied on written assessments, oral examinations (viva voce) and global ratings by clinical supervisors, despite recognition that these assessment methods were not completely suitable to the task of assessing the competence of (future) physicians. After the Second World War, assessment practices changed considerably (Norcini and McKinley, 2007; Van der Vleuten, 1996). Drivers of change in assessment practices were (amongst others):

- Advances in psychometrics, the 'basic science of assessment', and technology, enabling considerable gains in quality and efficiency of assessment;
- Quests for 'objectivity' as a direct reflection not only of the behaviourist tradition in education at that time, but also of the search for 'objective, reproducible and thus fair' assessment data (i.e. reduction of subjectivity in assessment and improvement of measurement characteristics);
- Problems of logistics and resources in assessment development and administration, resulting from exponential growth in student numbers;

- Increasing public pressure for accountability, requiring medical schools, the profession and licensing bodies to provide evidence of assessment quality, ensuring that newly qualified doctors are indeed 'competent'.

Development of new assessment methods was based on the conception of professional competence as an aggregate of different components or attributes that were relatively distinct from each other, such as knowledge, procedural skills, communication skills, and professionalism (Van der Vleuten, 1996). Development of competence was then to be inferred from growth or development in each of the constituents. As a consequence, assessment of professional competence focused on measurement of the separate components or building blocks of professional competence and trainees were assumed to have achieved minimally acceptable levels of competence after having passed all relevant tests.

At the same time, assessment developments were characterized by attempts to approximate 'real life professional practice' – in order to enhance test validity while maintaining standardized assessment conditions. Educators increasingly focused on assessment techniques requiring students to "show how" they do things, instead of assessing whether students "know" things or "know how" to do things (Miller, 1990). The development of live simulations of clinical situations, aimed at assessing trainees' actual performance, marked significant progress in assessment of professional competence. The so-called performance-based tests (also known as OSCEs or SP-based tests) are now being part of high-stakes examinations in medical education all over the world.

Performance-based assessments: lessons learned

Performance-based assessments are well-researched and many valuable lessons have been learned from what appear to be consistent research findings.

Performance-based tests allow for direct observation of trainee performance and thus provide unique opportunities for feedback, coaching and debriefing, facilitating the development of clinical skills (Yudkowsky, 2009). Similar to other forms of simulation, performance-based tests enable medical teachers and trainees to make sure that a minimal level of competency is reached before trainees are allowed to deal with real patients. Moreover, scores on performance-based tests appear to predict future practice performance (Tamblyn et al., 2007; Wenghofer et al., 2009). Increasing concerns about patient safety and quality of health care are therefore likely to consolidate the use of performance-based testing in (high stakes) assessment programmes in medical education.

Studies on performance-based assessments consistently confirm one of the most important findings in all measurements of clinical competence, namely the task- or case-specificity of performance. Correlations between scores on different tasks or items in any test are low, limiting the extent to which results from small samples of tasks can be generalized to broader domains. As a consequence, long testing times and broad sampling from the task domain are necessary to produce adequate reproducibility. Although consistency of scoring is another potential problem, it and other sources of score variability seem to be less important, or are easier to control with careful training and (even minimal) standardization and structuring of the assessment (Van der Vleuten, 1996). Second, development of performance assessments has proven to be difficult and resource-intensive. Logistics in test administration are often complex and may pose problems of comparability and feasibility (Swanson et al., 1995; Reznick and Rajaratanam, 2000; Yudkowsky, 2009).

Major concerns in the use of performance-based tests, therefore, are difficulties in test development and resource costs in terms of both money and time.

Given the importance of performance-based tests in assessment programmes in health professions education, and also given the fact that many educational institutions are confronted with declining budgets (e.g. Albanese, 2000), there is a need to find cost-effective ways for test development and administration, without compromising the quality of measurements.

Where are we going: workplace-based assessment revisited

In general, assessment practices until the beginning of the 21st century can be characterized by the development of a broad range of standardized and 'objectified' assessment methods ranging from written assessments and computer-based simulations to performance-based tests, each aimed at assessing different competency areas or skills, with the purpose of generating 'objective' assessment data. Decisions about which methods to use typically rest on psychometric indices, the quantifiable measures of reliability and validity being key in 'assessment of assessment'.

Assessment practices in medicine, however, are changing as they are in other professional and educational domains. Recent developments in assessment are stimulated by several factors:

- Professional competence can no longer be defined as a set of prescribed and well-defined attributes that have to be acquired. Not only is there a wealth of research evidence proving this approach to be incorrect (Van der Vleuten, 1996; Van der Vleuten et al., 2010), rapidly and continuously changing working environments will also demand competencies and skills that have not yet been defined (Segers et al., 2003). New approaches and definitions of competence emphasize the *integrated* conception of competence, i.e. the ability to handle complex and demanding tasks in the professional domain, integrating relevant cognitive, psychomotor and affective skills (Van der Vleuten and Schuwirth, 2005; Hager et al., 1994).
- The ultimate goal of professional education is to ensure that trainees are well prepared to handle tasks in the real world. Although performance-based tests significantly advanced assessment of competence and performance, research findings show that traditional tests do not accurately predict what trainees or professionals actually *do* in real life practice (e.g. Rethans et al., 1991). We therefore need additional methods to assess *habitual* performance in day-to-day professional practice (Epstein and Hundert, 2002).
- There is an increasing demand for ways of monitoring, measuring and maintaining competence and performance of *practicing* physicians (Cunnington and Southgate, 2002). Highly publicized failures of medical performance, declining public confidence in the medical profession and professional self-regulation, increasing demands for physician accountability, concerns about patient safety and improvement of quality of care have resulted in a search for robust procedures to 'measure' and maintain physicians' competence and performance throughout their professional careers (Norcini, 2005; Crossley et al., 2002; Davies, 2005).

- Assessment increasingly emphasizes formative assessment over summative assessment. Professional competence is to be developed and maintained through deliberate practice and reflection on experience (Ericsson, 2004). Competence is a habit of lifelong learning rather than an 'once-in-a-lifetime achievement' (Epstein, 2007, Segers et al., 2003). Continuous evaluation of as well as meaningful, focused feedback on performance are therefore key to professional competence development.

In terms of assessment practices these developments imply a need for greater authenticity in assessments, a shift from isolated tests to multiple assessments in 'real life' contexts and integration of assessment in learning and working processes. As a result growing emphasis on workplace-based assessment (WBA) has become an important component of current changes in assessment practices (Davies, 2005; Norcini, 2005).

Clear advantages of WBA include frequent and continuous opportunities for direct observation of and 'real-time' feedback on performance in handling authentic professional tasks – enabling competence development. Intuitively, WBA also seems to be the most valid way of assessing professional competence, as it is the best way of collecting multiple data on trainees' and professionals' habitual performance in day-to-day practice. From a theoretical perspective, workplace-based assessment strategies appear to be able to pre-eminently serve both formative and summative assessment purposes. Research findings, however, raise serious concerns about the utility and quality of WBA.

Firstly, assessment tasks in the real world setting are unpredictable and inherently unstandardized and they will not be equivalent over different administrations. From the psychometric perspective, this poses serious threats to reliability, validity (and fairness) of assessment.

Secondly, as professional judgment is inherent in WBA, serious concerns are raised about the subjectivity of assessments. Raters are generally considered to be major sources of measurement error (Albanese, 2000; Downing, 2005). Performance ratings are considered to be unacceptably biased, suffering from halo and leniency effects, and intra- and interrater reliability of performance ratings are often found to be substandard (Kreiter and Ferguson, 2001; Van Barneveld, 2005; Turnbull and Van Barneveld, 2002; Albanese, 2000). Lack of rater training, lack of structure in assessment mechanics (rating scale format, assessment criteria, for instance) and idiosyncrasy of rater performance theories are identified as potential sources of problems in WBA. Efforts to improve the quality of WBA therefore have focused on rater training and standardization of assessment procedures – with mixed success thus far (Williams et al., 2003; Lurie et al., 2009; Green and Holmboe, 2010).

A more serious finding, however, seems to be that formal observation of trainees in the clinical setting is rare (Day et al., 1990; Holmboe, 2004). The resulting lack of meaningful and useful feedback is a barrier to effective development of competence and expertise. Lack of direct observation also, inevitably, results in lack of documentation of performance. Lack of documentation is considered to be one of the major reasons for inaccuracy in performance ratings and failure to fail (DeNisi et al., 1989; Dudek et al., 2005). Frequent, continuous documentation of performance evaluations could therefore not only provide robust feedback for competence development, it could also decrease generosity error and improve summative decision making. Others, however, argue that WBA is to primarily serve a developmental function (Norcini and Burch, 2007; McGaghie, 2009), and that the use of formative assessments for summative purposes may impact negatively on trainee-supervisor (rater) relationship as well as on the utility of assessment results (Toohey et al., 1996; Hays and Wellard, 1998).

Obviously, given the increased significance of WBA in health professions education, there is a need to improve assessment practices. Key issues and research questions that emerge from findings as described above are:

- How does structuring or standardization of WBA improve assessment quality, for purposes of feedback and summative decision making?
- How can we optimize the use of performance assessment results? Can formative assessments be used for summative purposes? What are preconditions for effectiveness and efficiency?
- How can we improve performance ratings and make better use of professional judgments? What influences rater behaviours, and what explains differences in rater behaviours? What are reasons for persistence in rater behaviours, despite rater training? What are the processes underlying judgment and decision making?

Thesis outline

The research presented in this thesis addresses some of the issues and questions that are raised with respect to performance-based tests and workplace-based assessment.
Chapter 2 is a prelude to chapter 3 describing efforts in midwifery education to develop a performance-based test, aiming to assess professional competence in student-midwives. In midwifery education, as in other professional domains, dissatisfaction with conventional assessment methods led to the search for more authentic assessments, that would enable assessment of performance in handling realistic and complex tasks. Based on research findings in medical education and adopting an integrated conception of professional competence, a 6-station, 3-hour OSCE-based examination was developed, focusing on risk assessment and management in midwifery. The test was implemented as part of the examinations during the final stages of the training programme. Difficulties in test development and administration are described as well as perceptions of the usefulness and acceptability for comprehensive assessment of competence.

Chapter 3 addresses some of the major concerns regarding performance-based testing, i.e. limited generalizability of test results and concerns about test efficiency. Research findings in medical education show that the reproducibility of scores on performance-based tests is often problematic, because of limited sampling from the broad domain of professional tasks. The purpose of our study was to investigate whether reproducibility problems would be reduced by narrowing the task domain. We compared the reproducibility of OSCE-based scores in the (narrow) domain of midwifery to that of OSCEs in the (broad) medical domain. We also investigated effects of rating scale format (checklists versus global rating scales) and adoption of a mastery-oriented test perspective on the reproducibility of test scores in order to explore possibilities to increase test efficiency.

Chapter 4 presents a study in which student perceptions of a (semi-)structured approach towards WBA are described. More specifically, we investigated the effects 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 (i.e. daily) documentation of performance using standardized checklists. The main purposes of the observational diaries were to support and guide student learning throughout the clerkship. Documented formative feedback and performance evaluations in the diary were also used qualitatively to support decision making about competence development and achievement. Frequent and continuous assessments of performance were thus used for both formative and summative purposes. Using focus group techniques, we explored student perceptions on the impact of the integrated assessment approach on student learning and supervisor-rater behaviour. We furthermore investigated how students perceived the usefulness of observational diaries for summative decision making, focusing on credibility and defensibility of assessment procedures, perceived fairness and preconditions for success.

Chapters 5-7 focus on professional judgment. Chapter 5 summarizes key findings in the literature on performance appraisal and assessment, drawing from research in various professional fields. We discuss raters' judgment and decision making processes from a social cognitive perspective. We argue that assessment outcomes are determined by raters' cognitive processes which are very similar to reasoning, judgment and decision making in, for instance, clinical reasoning. Furthermore, we focus on contextual factors that influence rater behaviours as well as on the relationship between performance theories and values and the practice of performance assessment. Based on insights from other disciplines, we propose an alternative approach towards performance assessment, taking a constructivist, social-psychological perspective.
Research presented in chapter 6 builds on the social-psychological, constructivist approach towards assessment as proposed in chapter 5. Using the theoretical frameworks of cognition-based assessment and expertise research, we explored the cognitive processes that underlie judgment and decision making by raters when observing performance in the clinical workplace. We specifically investigated if and how differences in rating experience influence raters' information processing. Using verbal protocol analysis we investigated how experienced and non-experienced raters select and use observational data to arrive at judgments and decisions about trainees' performance in real-life patient encounters.
In chapter 7 we argue that assessment of performance in work settings should be seen as a specific application of social perception to specific purposes and that, consequently, much of raters' behaviours can considered to be rooted in social perception phenomena. Findings from research in social cognition and social perception indicate that most people, when judging others, make use of pre-existing knowledge structures, so-called *schemas.* The study presented in chapter 7 explored the use of schemas by physician raters during assessment of trainee performance in single patient encounters, using the theoretical frameworks of social perception. More specifically we explored the implicit 'role schemas' or performance theories of physician-

raters in general practice; the use of task- or ‘event’-specific performance schemas and the formation of person schemas during observation and assessment of performance. We furthermore explored if and how the use of performance schemas is influenced by different levels of rater expertise.

Finally, in chapter 8 the findings from our studies are summarized and discussed. Implications for assessment of educational and professional practice as well as challenges for further research are described.

Note
This thesis is a collection of related articles. Since every chapter was written to be read on its own, repetition and overlap across chapters are inevitable.

References

Abrahamson, S. (2000). Medical education: The testing of a hypothesis. In: L.H. Distlehorst, G.L. Dunnington & J.R. Folse (Eds.), *Teaching and Learning in Medical and Surgical Education: Lessons Learned for the 21st Century* (pp. 1-15). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Albanese, M.A. (2000). Challenges in using rater judgments in medical education. *Journal of Evaluation in Medical Practice, 6*(3), 305-319.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Bolk, J.H. (2000). Medisch Onderwijs in de 21e eeuw: Anders dan voorheen? Inauguration address, University of Leiden.

Brooks, M.A. (2009). Medical education and the tyranny of competency. *Perspectives in Biology and Medicine, 52(*1), 90-102.

Burack, J.H., Irby, D.M., Carline, J.D., Root, R.K. & Larson, E.B. (1999). Teaching compassion and respect: Attendings' responses to problematic behaviour. *J Gen Intern Med, 14*(1), 49-55.

Carraccio, C., Wolfsthal, S.D., Englander, R., Ferentz, K. & Martin, C. (2002). Shifting paradigms: From Flexner to competencies. *Academic Medicine, 77*(5), 361-367.

Choudhry, N.K., Fletcher, R.H. & Soumerai, S.B. (2005). Systematic review: The relationship between clinical experience and quality of health care. *Ann Intern Med, 142*, 260-273.

Crossley, J., Humphris, G. & Jolly, B. (2002). Assessing health professionals. *Medical Education, 36*, 800-804.

Cunnington, J. & Southgate, L. (2002). Relicensure, recertification and practice-based assessment. In: G.R. Norman, C.P.M. van der Vleuten & D.I. Newble (Eds.), *International Handbook of Research in Medical Education* (pp. 883-912). Dordrecht: Kluwer Academic Publishers.

Dauphinee, W.L. (2002). Licensure and certification. In: G.R. Norman, C.P.M. van der Vleuten & D.I. Newble (Eds.), *International Handbook of Research in Medical Education* (pp. 835-882). Dordrecht: Kluwer Academic Publishers.

Day, S.C., Grosso, L.G., Norcini, J.J., Blank, L.L., Swanson, D.B. & Horne, M.H. (1990). Residents' perceptions of evaluation procedures used by their training program. *J Gen Intern Med, 5*, 421-426.

Davies, H. (2005). Work based assessment. *BMJ Career Focus, 331*, 88-89.

DeNisi, A.S., Robbins, T. & Cafferty, T.P. (1989). Organization of information used for performance appraisals: Role of diary-keeping. *Journal of Applied Psychology, 74*(1), 124–129.

Donaldson, L.J. (1994). Doctors with problems in an NHS workforce. *BMJ, 308*, 1277-1282.

Downing, S.M. (2005). Threats to the validity of clinical teaching assessments: What about rater error? *Medical Education, 39*(4), 353-355.

Dudek, N.L., Marks, M.B. & Regehr, G. (2005). Failure to fail: The perspectives of clinical supervisors. *Academic Medicine, 80*(10), S84-87.

Ericsson, K.A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 79*(10), S70-S81.

Epstein, R.M. & Hundert, E.M. (2002). Defining and assessing professional competence. *JAMA*, *287*(2), 226-235.

Epstein, R.M. (2007). Assessment in medical education. *New England Journal of Medicine 356,* 387-396.

Frederiksen, N. (1984). The real test bias. *American Psychologist, 39*(3), 193-202.

Ginsburg, S., Regehr, G., Hatala, R., McNaughton, N., Frohna, A., Hodges, B., Lingard, L & Stern, D. (2000). Context, conflict, and resolution: A new conceptual framework for evaluating professionalism. *Academic Medicine, 75*(10), S6-11.

Green, M. & Holmboe, E. (2010). The ACGME toolbox: Half empty or half full? *Academic Medicine, 85*(5), 787-90.

Hager, P., Gonczi, A. & Athanasou, J. (1994). General issues about assessment of competence. *Assessment & Evaluation in Higher Education*, *19*(1), 3-16.

Hauer, K.E., Teherani, A. Kerr, K.M., Irby, D.M. & O'Sullivan, P.S. (2009). Consequences within medical schools for students with poor performance on a medical school standardized patient comprehensive assessment. *Academic Medicine, 84*(5), 663-668.

Hawe, E. (2003). It's pretty difficult to fail: The reluctance of lecturers to award a failing grade. *Assessment and Evaluation in Higher Education, 28*(4), 371–382.

Hayes, C.W., Rhee, A., Detsky, M.E., Leblanc, V.R. & Wax, R.S. (2007). Residents feel unprepared and unsupervised as leaders of cardiac arrest teams in teaching hospitals: A survey of internal medicine residents. *Crit Care Med, 35*(7), 1781-2.

Hays, R. & Wellard, R. (1998). In-training assessment in postgraduate training for general practice. *Medical Education, 32*, 507–513.

Holmboe, E.S. (2004). Faculty and the observation of trainees' clinical skills: Problems and opportunities. *Academic Medicine*, *79*, 16-22.

Howley, L.D. (2004). Performance assessment in medical education: Where we've been and where we're going. *Eval Health Prof, 27*, 285-303.

Ilott, I. & Murphy, R. (1997). Feelings and failing in professional training: The assessor's dilemma. *Assessment & Evaluation in Higher Education, 22*(3), 307-316.

Iobst, W.F., Sherbing, J., ten Cate, O., Richardson, D.L., Dath, D., Swing, S.R., Harris, P., Mungroo, R., Holmboe, E.S. & Frank, J.R. (2010). Competency-based medical education in postgraduate medical education. *Medical Teacher, 32*, 651-656.

Klass, D. (2007). Assessing doctors at work - Progress and challenges. *N Engl J Med*, *365*, 414-415.

Kogan, J.R., Holmboe, E.S. & Hauer, K.E. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. *JAMA, 302*(12), 1316-1326.

Kreiter, C.D. & Ferguson, K.J. (2001). Examining the generalizability of ratings across clerkships using a clinical evaluation form. *Evaluation & The Health Professions, 24*, 36-46.

Leape, L.L. & Fromson, J.A. (2006). Problem doctors: Is there a system-level solution? *Ann Intern Med, 144*, 107-115.

Littlefield, J.H., DaRosa, D.A., Anderson, K.D., Bell, R.M., Nicholas, G.G. & Wolfson, P.J. (1991). Assessing performance in clerkships: Accuracy of surgery clerkship performance raters. *Academic Medicine, 66*(9), S16–S18.

Lurie, S.J., Mooney, C.J. & Lyness, J.M. (2009). Measurement of the general competencies of the Accreditation Council for Graduate Medical Education: A systematic review. *Academic Medicine, 84,* 301-309.

Lysaght, R.M. & Altschuld, J.W. (2000). Beyond initial certification: The assessment and maintenance of competency in professions. *Evaluation and Program Planning, 23*(1), 95-104.

Mangione, S. & Nieman, L.Z. (1997). Cardiac auscultatory skills of internal medicine and family practice trainees. A comparison of diagnostic proficiency. *JAMA, 278*(9), 717-722.

McGaghie, W.C., Butter, J. & Kaye, M. (2009). Observational assessment. In: S.M. Downing & R. Yudkowsky (Eds.), *Assessment in Health Professions Education* (pp. 185-216). New York, NY: Routledge.

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

Norcini, J. & Burch, V. (2007). Workplace-based assessment as an educational tool. AMEE Guide No. 31. *Medical Teacher, 29*(9), 855-71.

Norcini, J.J. & McKinley, D.W. (2007). Assessment methods in medical education. *Teaching and Teacher Education, 23*, 239-250.

Norcini, J.J. (2005). Current perspectives in assessment: The assessment of performance at work. *Medical Education, 39*, 880-889.

Papadakis, M.A., Teherani, A., Banach, M.A., Knettler, T.R., Rattner, S.L., Stern, D.T., Veloski, J.J. & Hodgson, C.S. (2005). Disciplinary action by medical boards and prior behavior in medical school. *N Engl J Med, 353*, 2673-82.

Perez, J.A. & Greer, S. (2009). Correlation of United Stated medical licensing examination and internal medicine in-training examination performance. *Advances in Health Professions Education, 14,* 753-758.

Reilly, B.M. (2003). Physical examination in the care of medical inpatients: An observatory study. *Lancet, 362*(9390), 1100-1105.

Rethans, J.J., Sturmans, F., Drop, R., Van der Vleuten, C. & Hobus, P. (1991). Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. *BMJ, 303*, 1377-1380.

Reznick, R.K. & Rajaratanam, K. (2000). Performance-based assessment. In L.H. Distlehorst, G.L. Dunnington & J.R. Folse (Eds.), *Teaching and Learning in Medical and Surgical Education: Lessons Learned for the 21st Century* (pp. 237–244). Mahwah: Lawrence Erlbaum Associates.

Segers, M., Dochy, F. & Cascallar, E. (2003). The era of assessment engineering: Changing perspectives on teaching and learning and the role of new modes of assessment. In: M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising New Modes of Assessment: In Search of Qualities and Standards* (pp. 1-12). Dordrecht: Kluwer Academic Publishers.

Schwind, C.J., Williams, R.G., Boehler, M.L. & Dunnington. G.L. (2004). Do individual attending post-rotation performance ratings detect resident clinical performance? *Academic Medicine*, *79*, 453-457.

Struyven, K., Dochy, F. & Janssens, S. (2003). Students' perceptions about new modes of assessment in higher education: A review. In: M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising New Modes of Assessment: In Search of Qualities and Standards* (pp. 171-223). Dordrecht: Kluwer Academic Publishers.

Swanson, D.B., Norman, G.R. & Linn, R.L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher, 24*(5), 5-11+35.

Tamblyn, R., Abrahamowicz, M., Dauphinee, W.D., Hanley, J.A., Norcini, J., Girard, N., Grand'Maison, P. & Brailovsky, C. (2002). Association between licensure examination scores and practice in primary care. *JAMA, 288*, 3019-3026.

Tamblyn, R., Abrahamowicz, M., Dauphinee, M., Wenghofer, E., Jacques, A., Klass, D., Smee, S., Blackmore, D., Winslade, N., Girard, N., Du Berger, R., Bartman, I., Buckeridge, D.L. & Hanley, J.A. (2007). Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA, 298*(9), 993-1001.

Toohey, S. Ryan, G. & Hughes, C. (1996) Assessing the practicum. *Assessment and Evaluation in Higher Education, 21*(3), 215–227.

Van Barneveld, C. (2005). The dependability of medical students' performance ratings as documented on in-training evaluations. *Academic Medicine*, *80*(3), 309-312.

Van der Hem-Stokroos, H.H., Scherpbier, A.J.J.A., Van der Vleuten, C.P.M., de Vries, H. & Haarman, H.J. (2001). How effective is a clerkship as a learning environment? *Medical Teacher, 23*, 599-604.

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

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

Van der Vleuten, C.P.M., Schuwirth, L.W.T., Scheele, F., Driessen, E.W. & Hodges, B. (2010). The assessment of professional competence: Building blocks for theory development. *Best Practices & Research Clinical Obstetrics and Gynaecology.* Doi: 10.1016/j.bpobgyn.2010.04.001.

Wenghofer, E., Klass, D., Abrahamowicz, M., Dauphinee, D., Jacques, A., Smee, S., Blackmore, D., Winslade, N., Reidel, K., Bartman, I. & Tamblyn, R. (2009). Doctor scores on national qualifying examinations predict quality of care in future practice*. Medical Education,* 43, 1166-1173.

Wilkerson, L. & Lee, M. (2003). Assessing physical examination skills of senior medical students: Knowing how versus knowing when. *Academic Medicine, 78*(10), S30-S32.

Williams, R.G. & Dunnington, G.L. (2004). Prognostic Value of Resident Clinical Performance Ratings. *J Am Coll Surg, 199*, 620-27.

Williams, R.G., Dunnington, G.L. & Klamen, D.L. (2005). Forecasting residents' performance – Partly cloudy. *Academic Medicine, 80*(5), 415-422.

Williams, R.G., Klamen, D.A. & McGaghie, W.C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine, 15*(4), 270-292.

Yudkowsky, R. (2009). Performance tests. In: S.M. Downing & R. Yudkowsky (Eds.), *Assessment in Health Professions Education* (pp. 217-244). New York, NY: Routledge.

CHAPTER **2**

Assessment of competence in midwifery education using OSCE technology

Published in British Journal of Midwifery 2001, vol 9, No 3: 156-161

(Editor's title: Objective assessment is needed to ensure competence)

Marjan J.B. Govaerts
Lambert W.T. Schuwirth
Annerien Pin
Mieke E.J. Clement
Cees P.M. van der Vleuten

Abstract

Over the last decades, assessment in medical education has changed considerably, partly as a result of increasing dissatisfaction with current professional competence evaluation. New strategies for the assessment of professional competence show an increasing emphasis on performance-based assessment (also known as "authentic" assessment). Changes have been prompted by the awareness of the enormous influence of assessment on student learning.

This paper will describe the development and the use of a performance-based assessment test at Kerkrade Midwifery School (the Netherlands)[1]. The OSCE-based Midwifery Competence Test (MCT) was designed to assess competence in risk assessment and management in midwifery care, during the final stages of the educational programme.

First experiences with the MCT suggest that the test is a powerful tool in the guidance of student learning. Implementation of the MCT furthermore stimulated reflections on and discussions about the essence of midwifery competence, thus contributing to improvement of the curriculum, teaching and instruction. Our experiences suggest that, although several aspects of the test need further attention, the MCT is a useful and acceptable tool for evaluating competence of student-midwives.

[1] Kerkrade Midwifery School is now known as the Faculty of Midwifery Education and Studies, Zuyd University (Academie Verloskunde Maastricht)

Introduction

In society today many professions are confronted with rapid changes in professional practice and shifting concepts of professional roles, tasks and responsibilities. These changes necessarily force professional schools to revise the concept of professional competence and to adapt curricula to keep up with new expectations (Curry and Wergin, 1993). Societal and political forces, information technology, and increasing demands for professional accountability have implications for midwifery practice and education in midwifery schools (Bonsel and van der Maas, 1994; Avery, 1999; Sinclair and Gardner, 1999; Thompson, 1999). Concern with educating ‘competent’ midwives will not only make a redesign of curricula inevitable, but will also emphasize the need for evaluation of common assessment procedures and development of new, appropriate assessment techniques.

In many professional training programs there has been a growing dissatisfaction with conventional assessment instruments (McGaghie, 1991; Hager and Butler, 1996). This has been for several reasons. First, professional competence evaluation is often biased toward assessment of knowledge rather than the assessment of professional performance capabilities. Many assessment techniques rely on written tests. Although these tests can be seen as reliable and objective assessment measures, they fail to reflect the complexity of real professional practice and are lacking in direct validity. Similar drawbacks go for assessment of specific skills in isolation, out of the professional context. Second, evaluation of clinical skills is often done by experts using indirect and unstructured methods. Most clinical evaluations rely on subjective judgements, second-hand information and global impressions of competence level, which make reliability questionable. Third, many measures of competence sample only a small range of skills and professional contexts. Complex professional performance not only involves knowledge or technical skills, but also professional judgement, personal values and attitudes. Finally, educational innovations such as problem-based learning and competency-based education challenge the value of traditional assessment practices. Although the relationship between assessment and learning behaviour of students is well established (e.g. Frederiksen, 1984), many assessment techniques do not reflect new and changing educational objectives.

In response to these problems in competence measurement, new developments in assessment strategies show an increasing emphasis on assessment methods involving complex practices and procedures in a professional context, i.e. the context in which they are actually used. These methods are often referred to as “authentic” or performance-based assessments. Performance-based assessments have been developed in medical education, nursing education, education of teachers, power plant operators, managers, etc. (e.g. McKnight et al., 1987; Van der Vleuten and Swanson, 1990; McGaghie, 1993). The present article describes the development and the use of a performance-based assessment test in midwifery education at Kerkrade Midwifery School in the Netherlands.

Competence

Defining competence is necessary in order to be clear about what it is one wants to assess. On reviewing the literature, it appears that several very different approaches to defining and assessing professional competence are used (Norman, 1985; Girot, 1993; Hager and Gonczi, 1996). The same applies to competence of midwives (e.g. Brogan, 1999; Cairns and Thomson, 1999; Hindley, 1999).

For the purpose of this article, competence will be defined as the ability of a person to handle the specific situations that arise in his or her area of practice effectively (Kane, 1992; Hager and Gonczi, 1996). According to this definition, competence is an integrated concept in terms of knowledge, skills, judgement and attitudes that are required for the competent performance of professional tasks. There are two major components in this concept of competence. One is the domain of possible professional tasks and the other includes the knowledge, skills, attitudes and judgement that the professional is expected to use in handling these tasks.

Midwifery and midwifery competence

The Dutch health care system recognises midwives who work in primary health care as autonomous care providers with their own powers and responsibilities. Midwives work with the basic assumption that pregnancy, delivery and the puerperium are physiological, natural processes. Midwives will monitor every pregnancy and check for pathological deviations from the physiological course. The midwife in primary health care is responsible for the detection and assessment of possible medical and non-medical risk factors and determining appropriate obstetric management (Crébas, 1989).

This description of professional duties of midwives implies capabilities to:

1. Acquire relevant information (history taking, physical examination, laboratory investigations, etc.);
2. Interpret the significance of the information obtained in assessing possible risk factors and making a diagnosis;
3. Determine the appropriate client management using risk assessment;
4. Apply relevant knowledge, skills, professional attitude and judgement in order to provide effective and efficient care;
5. Communicate findings and management decisions to the woman and/or other (obstetric) health-care providers.

This description of midwifery competence resembles descriptions of clinical competence in other health-care professions like medicine and nursing (Norman, 1985; McKnight et al., 1987).

Assessment of professional competence

Several innovative techniques have been developed for assessment of clinical competence. Developments are characterized by continuous attempts to approximate the real professional world as much as possible, while maintaining standardized test-taking conditions (Van der Vleuten, 1996). One of the most successful innovations is the introduction of ‘multiple-station examinations’, mostly known as OSCEs (Objective Structured Clinical Examinations) and SP-based tests (Standardized Patient-based tests). During a multiple-station examination the student rotates around a series of standardized simulated professional tasks, called stations. Clinical reality may be simulated by the use of standardized (real or simulated) patients. At each station, students' performances are recorded by trained faculty staff examiners or by trained standardized patients using pre-coded checklists and rating forms. Use of multiple-station examinations enables assessment of broad areas of clinical competence, combining standardization of test items with multiple direct observations of student performance. Rich medical literature referring to multiple-station examinations indicates that they appear to be a method of assessing competence that is objective in nature, fulfils the criteria of validity, reliability and practicality (Van der Vleuten, 1996; Ladyshewsky, 1999).

The Midwifery Competence Test

At Kerkrade Midwifery School (KMS), growing concern with the assessment procedures used, especially in the final examinations, initiated the development of more authentic competence evaluation. A multiple-station examination was implemented, designed to assess student performance in specific task situations in midwifery care.

Test design

The Midwifery Competence Test (MCT) consists of 6 different stations. Station length is 30 minutes each, resulting in a total test length of three hours. The test content reflects educational objectives with regard to the following task domains within the scope of midwifery practice:

1. Tasks to be performed during pregnancy (antenatal period);
2. Tasks to be performed during labour and delivery (intranatal period);
3. Tasks to be performed after delivery (postnatal period).

For each of these task domains, two stations are included in the test. At each station a midwifery care problem is presented to the student (e.g. breech presentation, umbilical cord prolapse or low haemoglobin). At *each* station students are evaluated for their ability to obtain relevant information from history taking and physical examination, application of knowledge to the problems presented, interpretation of findings, performance of technical skills and the ability to integrate all of these facets in risk assessment and management decision making. Competence of interpersonal skills is assessed at one or more stations, as an integrative part of client management skills. Diagnostic and therapeutic skills are performed on obstetric models (mannequins). In several stations standardized patients are used, especially in stations testing interpersonal skills. At each station two examiners rate student performance: one being a staff member, the other a practising midwife. Examiners use detailed scoring keys and pre-coded checklists for scoring performance (Tables 2.1 and 2.2).

Test preparation and administration

During preparation of each test, several meetings are planned for briefing and instructing the examiners and the simulated patients. Before actual use of a station in the MCT a pilot (trial) is conducted which provides feedback to the test developers. Feedback is used to re-evaluate and revise the case c.q. station.

Table 2.1 *Example of a scoring key and instructions to the examiner*

Instructions to the student (case description) Mrs. Van Waardenburg is pregnant and she is visiting you for the first time. Date of last menstruation: March 10, 1999. Mrs. Van Waardenburg is a Dutch female, 27 years old, G2, P1, A0. Eighteen months ago she gave birth to a son (3500 g birth weight, gestational age 39^{+4} weeks, delivery at home) There were no complications during pregnancy, delivery or postnatal period. Mrs. Van Waardenburg breast-fed her son for three months after delivery.	
1. Question **You draw some blood for blood tests. Which blood tests do you conduct now (or do you order?)**	
Scoring key	
Good	blood group; Rh type; HbsAg; Irregular antibodies (IEA); Syphilis screen; Haemoglobin (Rubella antibody titre and haematocrit may be ordered, but these tests are not commonly conducted in primary care)
Fair	student names 5 tests
Satisfactory	student names 4 tests
Unsatisfactory	student names less than 4 tests
Student receives form with lab results. **2a. Question** **Please interpret the results of the laboratory tests and explain your answer.**	
Scoring key	
Good	haemoglobin is pathologically low, so Mrs. Van Waardenburg suffers from anaemia. Since she is Dutch, iron deficiency anaemia will be the most probable diagnosis. Mrs. Van Waardenburg is Rh-negative. This might lead to complications from isoimmunisation.
Fair	diagnosis correct, without explanation
Satisfactory	only names low haemoglobin/anaemia or rhesus-immunization
Unsatisfactory	all other answers
2b. Question **Determine appropriate obstetric management, based on your interpretation of lab results.**	
Scoring key	
Good	a. prescription of iron supplementation b. start iron intake as soon as possible c. initial provision of iron salts for 6 weeks d. repeat Hb after 6 weeks e. check if blood group of first child was determined and if D-positive, check if the woman received anti-D immunoglobulin f. maternal antibody determination by 30 weeks gestation g. determination of rhesus status of child immediately after birth (umbilical cord blood)
Fair	names a., b., f., and g.
Satisfactory	names a. and f.
Unsatisfactory	all other answers.
Examiner may add that the question involves obstetric care during pregnancy, delivery and postnatal period.	

The preparation and administration of the MCTs falls within the responsibility of a group of faculty members, the MCT Committee, consisting of the clinical skills co-ordinator and experienced midwives. The MCT Committee selects problems to be included in the test and determines station subject matter, using a test blueprint. The MCT Committee is also responsible for reviewing all stations for content, relevance and level of difficulty. Stations are developed by faculty members (midwives mostly).

All students in the final stages of their training program (third and fourth year) participate in the MCT. For each MCT two separate series of stations are prepared, one for each year group.

Table 2.2 *Example of a checklist*

Midwifery Competence Test
Station number
Date
Student number
Name student
Examiners

ANTENATAL PERIOD	Good	Fair	Satisfactory	Unsatisfactory	Wrong	Not done
History/physical examination/laboratory investigations						
1. Which blood tests do you conduct or do you order?	0	0	0	0	0	0
6. Which additional information do you need in order to be able to determine appropriate obstetric care?	0	0	0	0	0	0
Risk assessment and obstetric management						
2a. Give an interpretation of the lab results and explain your answer.	0	0	0	0	0	0
2b. Determine obstetric management and formulate a patient management plan.	0	0	0	0	0	0
7. What would be appropriate obstetric management at this moment, considering available information? Explain your answer.	0	0	0	0	0	0
Theoretical knowledge base						
4. Name most common causes of anaemia during pregnancy.	0	0	0	0	0	0
5. Name possible consequences of isoimmunisation.	0	0	0	0	0	0
Technical skills						
3. Write out a prescription (iron treatment) according to guidelines.	0	0	0	0	0	0

Student number
Student name

Midwifery Competence Test
Year

	Antenatal Station 1 % score	Antenatal Station 2 % score	Intranatal Station 4 % score	Intranatal Station 5 % score
1. History/physical/lab. investigations	100	33	85	70
2. Risk assessment and management	100	76	82	73
3. Theoretical knowledge base	100	100	100	100
4. Technical skills	100	100	100	100
Total % score	100	71	88	81
Student score on task domain	85		85	
Mean % score reference group (standard deviation) on task domain	80 (10)		70 (5)	
Result	pass		pass	

Note Students will receive a ‘pass’ on a task domain if the students’ percentage score on the domain equals or exceeds the mean percentage score of the reference group (year group) minus 1.0 standard deviation.
Minimally acceptable task domain score is set at 55%; students with a task domain score less than 55% will fail the examination. Students with a task domain score that equals or exceeds 70% will receive a ‘pass’ on the domain.

Figure 2.1 Example of feedback form for student

The students rotate around the series of 6 stations. The test is extended over two consecutive days, students visiting three stations each day.

Pass / fail decisions

Faculty decided that students need to show competence in any of the three task domains within midwifery practice mentioned above. Each case in the MCT receives a single score. Scoring on a case is calculated as a percentage of possible points based on all elements in the case. For each task domain a pass/fail decision is determined by averaging the scores on the two cases representing the domain. A norm-referenced standard setting method is applied, using the mean score minus one standard deviation to set the standard. However, students must show a minimum level of performance on all task domains to pass the MCT: for each task domain a minimum acceptable cut score is set at 55%. After test administration students receive written feedback about their performance levels (Figure 2.1).

Table 2.3 *Mean score (scale 1 = fully disagree, 5 = fully agree) and corresponding standard deviation (SD) per item in the questionnaire (relevant items), and for the overall judgment of utility of assessment strategies in assessment of midwifery competence (scale 1-10)*

	Students N = 72		Examiners N = 42	
	Mean (1-5)	SD	Mean (1-5)	SD
MCT content-procedures overall satisfactory	4.3	.8	4.7	.5
Station content relevant	4.1	.8	4.6	.6
MCT valuable in assessment of competence	3.6	1.0	4.3	.6
Station content representative	3.6	1.0	4.6	.6
Examiner behaviour correct	4.2	.7	Questions not in examiner questionnaire	
MCT important part of final examinations	4.0	.9		
MCT valuable part of educational programme (instructive)	3.9	1.1		
Overall judgment of utility (scale 1-10):				
Progress test (knowledge test)	7.1	1.2		
MCT	7.4	1.1		
In-training evaluations (work-based assessments during clerkships)	8.8	1.2		

Note Students' response rate is 94%; examiners' response rate is 91%

First experiences

Overall, first experiences with the MCT are encouraging. After test administration in 1999, structured questionnaires were used to explore students' and examiners' perceptions of and satisfaction with the MCT. The questionnaires consisted of statements to be responded to on a five-point Likert scale (1 = fully disagree; 5 = fully agree). The questionnaires were designed to elicit feedback on several test aspects and contained questions about the examination procedure; relevance of station content; adequacy of the instrument in competence measurement; examiner behaviour and student preparation. The students were also asked to give a mark out of 10 for the usefulness of three different assessment methods (including MCT) for assessment of professional competence. Finally, the questionnaire contained open questions that invited respondents to add comments and to provide suggestions for improvement.

Results in Table 2.3 show students' and examiners' responses to relevant items, indicating a fairly positive perception of the utility of the MCT. Both students and staff members see the MCT as a valuable tool in the assessment of professional competence of student-midwives. Station content is considered relevant and fairly representative for midwifery practice. It tests application of knowledge, skills and judgement necessary in a midwife-client encounter. In their comments on preparation students especially report positive effects on learning behaviour shifting from rote memorisation of facts towards critical thinking and application of knowledge in midwifery practice.

Students would typically comment that

> *"the MCT stimulated me to study. More importantly, it stimulated me to take a different approach: less memorizing, focusing more on clinical problems and how to deal with these patients."*
>
> or
>
> *"the MCT stimulated me to discuss clinical cases with my fellow-students or my supervisor, and to use my practical experiences as a starting point for learning."*

Staff feels that the MCT is an objective and valid means of assessment of midwifery competence. As opposed to typical OSCEs during which the competence to be tested is broken down into its various components, the MCT can be seen as a more comprehensive evaluation. At each station, it not only assesses students' knowledge and technical skills, but also abilities in entire client assessment, including problem solving, evaluation and management. Moreover, staff claim that MCT stimulates discussions about the essence of midwifery competence, refining curriculum content and improving instruction. The use of performance-based 'authentic' assessments can thus be a powerful tool in staff development and curriculum quality assurance.

Although first evaluations suggest that the MCT is a useful instrument, several aspects of the test deserve particular attention in the future. Students spend excessive amounts of time in preparation for the test, sometimes neglecting other learning activities (students reported an average of more than 80 hours study time). This can be accounted for by the fact that the MCT is part of a high stakes examination, but from students' comments it became clear that poor integration of the assessment with the rest of the curriculum was an additional explanation. Implementation of alternative assessments clearly requires attention to integration of assessment and instruction.

The standard-setting procedure for the MCT produces a failure rate that is relatively high (25-30% failure rate) if checked against other test results. Current conjunctive standard-setting procedures might therefore need re-evaluation. Test results furthermore show that average case scores vary considerably, reflecting large variations in case difficulty and therefore test difficulty. For instance, in year 4 average case scores ranged from 61% to 83%. These results illustrate the need for extensive quality assurance in test construction (e.g. review of cases and scoring keys by expert panels and case pilots prior to test administration) as a prerequisite for appropriate standard setting.

A final issue of concern in the continued use of the MCT is its high cost, including developmental costs, financial compensations for simulated patients and participating midwives, salary costs of faculty staff and support staff. A remedy might be collaboration of midwifery schools to jointly address the developmental costs associated with MCTs.

Conclusion

Recent research into competence assessment in health professions education has yielded new test procedures that may be applicable to midwifery education. The present article describes an attempt to transfer these insights to midwifery, and explores the MCT as an example of developing more accurate measures of competence in midwifery education. The first evaluation results and experiences from educational practice show MCT is a promising tool for evaluating competence in risk assessment and management. It appears to be a method of assessing competence which fulfils criteria of acceptability, validity and practicality. In addition, MCT has positive effects on student learning behaviour and stimulates curriculum improvement.

Several issues deserve more attention for the near future. First, students and staff argue that integration of the performance assessment with the curriculum needs attention. Second, standard setting is an issue in need of further research. Third, collaboration between schools in developing performance-based assessment techniques is to be encouraged.

References

Avery, M.D. (1999). Core competencies in nurse-midwifery education. In: *Book of Proceedings of the 25th Triennial Congress of the International Confederation of Midwives* (pp. 29-32). Medcom International, Philippines.

Bonsel, G.J. & Maas van der, P.J. (1994). *Aan de wieg van de toekomst. Scenario's voor de zorg rondom de menselijke voortplanting 1995-2010.* (At the beginning of the future. Sketches of (health) care with respect to human reproduction 1995-2010). Bohn, Stafleu, Van Loghum: Houten.

Brogan, K.A. (1999). Essential competencies for basic midwifery practice. In: *Book of Proceedings of the 25th Triennial Congress of the International Confederation of Midwives* (pp. 78-81). Medcom International, Philippines.

Cairns, R. & Thomson, A.M. (1999). The notion of competence as elicited in a study of 'Fitness for Purpose' on completion of a pre-registration (long) midwifery education course. In: *Book of Proceedings of the 25th Triennial Congress of the International Confederation of Midwives* (pp.89-94). Medcom International, Philippines.

Crébas, A. (1989). *Beroepsomschrijving Verloskundigen. Beroepsomschrijving opgesteld in opdracht van de Nederlandse Organisatie van Verloskundigen* (Description of the duties of a midwife, written by order of the Netherlands Organisation of Midwives). Bilthoven.

Curry, L., Wergin, J.F. & Associates (1993). *Educating professionals: Responding to new expectations for competence and accountability*. Jossey-Bass Inc. Publishers, San Francisco.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193-202.

Girot, E.A. (1993). Assessment of competence in clinical practice: A phenomenological approach. *Journal of Advanced Nursing, 18*, 114-119.

Hager, P. & Butler, J. (1996). Two models of educational assessment. *Assessment and Evaluation in Higher Education, 21*(4), 367-378.

Hager, P. & Gonczi, A. (1996). What is competence? *Medical Teacher, 18*(1), 15-18.

Hindley, C. (1999). The assessment of clinical competency on an undergraduate midwifery programme: Midwives and student experiences. In: *Book of Proceedings of the 25th Triennial Congress of the International Confederation of Midwives* (pp. 237-241). Medcom International, Philippines.

Kane, M.T. (1992). The assessment of professional competence. *Evaluation & the Health Professions, 15*(2), 163-182.

Ladyshewsky, R. (1999). Simulated patients and assessment. *Medical Teacher, 21*(3), 266-269.

McGaghie, W.C. (1991). Professional competence evaluation. *Educational Researcher, 20*(1), 3-9.

McGaghie, W.C. (1993). Evaluating competence for professional practice. In: L. Curry, J.F. Wergin and Associates (Eds.), *Educating professionals: Responding to new expectations for competence and accountability* (pp. 229-261). Jossey-Bass Inc., Publishers, San Francisco.

McKnight, J., Rideout, E., Brown, B. et al. (1987). The objective structured clinical examination: An alternative approach to assessing student clinical performance. *Journal of Nursing Education, 26*(1), 39-41.

Norman, G.R. (1985). Defining competence: A methodological review. In: Neufeld, V., Norman, G.R., (Eds.), *Assessing Clinical Competence* (pp. 15-33). New York: Springer.

Sinclair, M. & Gardner, J. (1999). High technology in midwifery practice. In: *Book of Proceedings of the 25th Triennial Congress of the International Confederation of Midwives* (pp. 495-499). Medcom International, Philippines.

Thompson, C. (1999). Preparing a competent and confident midwife-practitioner: An Australian perspective. In: *Book of Proceedings of the 25th Triennial Congress of the International Confederation of Midwives* (pp. 544-555). Medcom International, Philippines.

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

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

CHAPTER 3

Optimizing the reproducibility of a performance-based assessment test in midwifery education

Published in Advances in Health Sciences Education 2002,
Vol 7, No 2: 133-145

Marjan J.B. Govaerts
Cees P.M. van der Vleuten
Lambert W.T. Schuwirth

Abstract

Despite problems concerning generalizability of test results – largely due to limited sampling of the task domain – and questions about test efficiency, the popularity of OSCEs in medical education has motivated the use of similar assessment methods in other health care domains. Purpose of this study was to investigate reproducibility of scores on an OSCE-based test in the relatively narrow domain of midwifery, as compared to the broad medical domain. The influence of global rating scales and adoption of a mastery-oriented test perspective on reproducibility of test scores was investigated in order to explore possibilities to increase test efficiency.

A 3-hr, 6-station performance based test was administered to third and fourth year students at Kerkrade Midwifery School (the Netherlands). Students' performance was recorded using a station specific checklist and a global rating scale.

For the 3-hr OSCE the generalizability coefficient based on checklist scores is 0.48. Based on global ratings alone, the generalizability coefficient is 0.61. Adjusted dependability indices are 0.85 based on checklist scores and 0.95 for global ratings respectively. Results suggest that even for small domains the problem of case-specificity remains a major impediment to high stakes performance testing. Findings furthermore suggest that use of global rating scales and professional expert judgements in professional competence evaluation is to be preferred to task specific checklists for reasons of reproducibility and efficiency in test development and administration. Adoption of a mastery oriented test perspective may reduce testing time requirements even further.

Introduction

Qualification procedures, such as licensure and certification, have to assure that professionals are competent to perform the roles and tasks defining professional practice and can take on the responsibilities accorded to their professional status. Assessment systems used in high stakes competence evaluation should therefore enable valid and reliable judgements of readiness for practice (Eraut, 1994; Kane, 1994). Over the last decades, there has been a strong tendency to use ‘performance-based’ or ‘authentic’ assessments containing high fidelity simulations of complex professional practice, while at the same time maintaining standardised test taking conditions (Kane, 1992; Van der Vleuten, 1996).

In medical education one format especially, the Objective Structured Clinical Examination (OSCE) or multiple station examination, has gained enormous popularity. Multiple station examinations are received with enthusiasm for their apparent high face validity (authenticity) and presumed influence on learning and teaching (Newble and Jaeger, 1983; Newble, 1988). However, research findings also indicate that issues that were largely solved by more traditional ‘objective’ testing re-emerge as serious problems (Kane et al., 1999). First, testing time requirements limit task sampling from the extended range of tasks in the target domain and most performance assessments will necessarily include only a small number of performances. The problem of task specificity (low correlations between scores on different tasks in the test) limits the extent to which results from small samples of tasks can be generalised to broader domains. The task- or case-specificity phenomenon has proven to be the most serious concern in the use of performance based assessments (Swanson et al., 1987; Swanson and Norcini, 1989; Van der Vleuten and Swanson, 1990). Although consistency of scoring is another potential problem, this and other sources of score variability seem to be less important, or can be better controlled with careful training and (even minimal) standardisation and structuring of the assessment (Van der Vleuten, 1996). Second, development of performance assessments has proven to be difficult and resource-intensive. Test administration procedures are complex, posing problems of comparability and feasibility (Swanson et al., 1995; Reznick and Rajaratanam, 2000).

Major impediments to implementation of performance-based testing in high stakes examinations, therefore, remain lack of generalizability of test results, largely due to limited domain sampling, and efficiency problems.

Despite these problems, OSCEs have become widely used tools in assessment of clinical competence. The popularity of OSCEs in medical education has motivated the use of similar assessment procedures in other health care professions. An interesting question is whether problems identified in competence evaluation in medicine are as prominent in related, but different health care domains. This article will report on an OSCE in midwifery education. It will address the most threatening psychometric property of OSCEs and will investigate the reproducibility of scores derived from an OSCE based performance test in midwifery. We are unaware of publications reporting on the reproducibility of OSCEs in the midwifery domain. Midwifery is a professional domain closely related to medicine, but showing several distinctive characteristics that might influence reproducibility findings on performance based assessments. The domain of professional encounters defining midwifery practice is definitely of a narrower scope than the domain of medical practice and there appears to be closer similarity across tasks within the midwifery domain. Conceptual and technical skills required by midwives are clearly prescribed and the variety of contexts in which midwifery practice takes place seems to be less than the contextual variance in medicine. One might argue that the content-specificity problem is inherent to broadness of the task domain, thereby assuming that narrowing the domain would reduce reproducibility problems due to limited task sampling. Therefore, our first research question concerns the reproducibility of scores of an OSCE as a function of testing time in the narrow domain of midwifery and how it compares to OSCEs in the broad medical domain (e.g. Swanson and Norcini, 1989; Van der Vleuten and Swanson, 1990).

Most midwifery institutions are relatively small educational institutions. Apart from measurement characteristics, efficiency in test preparation and test administration is crucial in continued use of performance-based tests. Therefore, in order for the potential to save resources, two additional research questions were investigated. First, we investigated the influence of global rating scales. Recent research reports have indicated that global rating scales in OSCEs have better construct validity over commonly used task-specific checklists (Regehr et al., 1998; Hodges et al., 1999). Research results also show that global rating scales have better inter-case reliability, despite poorer inter-rater reliability, resulting in equal or even better reproducibility of overall test scores (Cunnington et al., 1997; Cohen et al., 1997; Rothman et al., 1997). Hence, our second research question concerned the differential reproducibility of scores between checklists and global ratings.

Finally, literature suggests (e.g., Norcini and Swanson, 1989; Van der Vleuten and Swanson, 1990) using a mastery-oriented test perspective on score interpretation to ameliorate the usual poor reproducibility of OSCEs. In a mastery-oriented perspective, the reproducibility of pass-fail decisions is estimated, rather than reproducibility of scores. Research findings show that using a mastery-oriented test perspective can considerably reduce testing time requirements, depending on the position of the pass-fail point (cut-off score) relative to the mean performance. Our final research question therefore addressed this issue by investigating the reproducibility of pass-fail decisions to be taken with the midwifery OSCE.

Context of the study

This study was conducted in the Netherlands. The organisation of the obstetric care in the Netherlands differs from the obstetric care organisation in surrounding countries. In the Netherlands, midwives have become specialists in physiological obstetrics, independent practitioners holding a key position in the Dutch obstetric care system (Crébas, 1990). In the Netherlands, a strict division is made in the field of obstetrics between physiology and pathology. Midwives will monitor every pregnancy and check for pathological deviations from the physiological course. The midwife in primary health care is responsible for the detection and assessment of possible medical and non-medical risk factors and determining of appropriate obstetric management. She is an autonomous care provider with her own powers and responsibilities.

The curriculum of midwifery schools is a 4-year training program encompassing theoretical and experiential learning. After finishing midwifery school, the midwife is authorised to independently practice midwifery/physiological obstetrics.

Method

Midwifery Competence Test

The Midwifery Competence Test (MCT) is a performance-based test designed to assess competence of student-midwives in the final stages of their training program. Details of development, administration and scoring of the test are reported elsewhere (Govaerts et al., 2001). A short description of important characteristics follows.

MCT is an OSCE-based multiple station examination format, consisting of six different cases (stations). Station length is 30 minutes for each station, resulting in total test length of three hours. The content of the MCT focuses on risk assessment and management in midwifery, reflecting exit objectives that represent the knowledge, skills, professional judgement and attitudes required of the students by the time of certification (end of year 4). All students in the final stages of their training program (third and fourth year students) participate in the MCT. For each year group a separate series of stations is prepared. At each station, students are evaluated for their ability to obtain relevant information from history taking and physical examination, application of knowledge to the problems presented, interpretation of findings, performance of technical skills and the ability to integrate all of these facets in risk assessment and management decision-making. Diagnostic and therapeutic skills are performed on obstetric models (manikins). In several stations standardised patients are used, especially in stations testing interpersonal skills. Where appropriate, examiners role-play clients for portions of the examination. At each station, student performance is rated by two examiners using a station specific checklist (consisting of 10-15 items per station). Examiners are trained in the use of the checklist during training sessions in which station content and scoring procedure are discussed following station try-outs. After a student's performance checklist scores from both examiners are always compared and checked against each other, resulting in one consensus score to be used for pass-fail decisions.

In the context of this study, examiners were also asked to give an overall rating of midwifery competence on a five-point global rating scale (1 = unacceptable, 2 = borderline performance, 3 = acceptable, 4 = good, 5 = excellent). Examiners received only a brief comment on the purpose and the use of the global rating scale. They were explicitly asked to use their professional expertise and judgement in evaluating the students' performance. In doing so they were urged to evaluate performance as if judging professional competence of fellow midwives, using all criteria considered essential for good professional practice, even if criteria were not mentioned on the checklist. Examiners were asked to independently complete global rating scales before completing the checklist, avoiding interaction about these individual assessments.

In this way, for each case two independent global rating scores were available for each student and one consensus checklist score. Unfortunately, the original independent checklist scores were not available for analysis.

Statistical analyses

The Midwifery Competence test (MCT) was administered to all third-year and fourth-year students (N = 160) in the classes of 1999 and 2000. In total four MCTs were used for analysis. Blueprints for the four test forms differed somewhat in content, but stations were similar in format and general intent. Because examiners not always completed global rating scales, only students for whom both checklist scores and global rating scores were available were included in the study.

For each station, one checklist score was available for each student, which was a percentage of possible points based on all elements in the case. For each station, one score on overall competence was calculated for each student by averaging the global ratings on midwifery competence. At the end of the examination for each student a final checklist-score and an overall midwifery competence score were computed, by computing the mean of the student's case checklist scores and of the global ratings on overall competence respectively across stations.

For the first research question generalizability theory was used. Using generalizability theory, different sources of variation affecting the measurement can be evaluated and their magnitude can be estimated by using analysis of variance and variance component estimation. The estimated variance components can be used then for decision studies to estimate reliabilities under different measurement conditions, for example by varying sample sizes of cases and raters. For each MCT, variance components for examinees, cases and residual effect were estimated using a random effects persons by stations design (PxS design). Variance components were then pooled across test administrations and years of training by averaging variance component estimates weighted by sample sizes. This provides a single and most stable estimate of variance components to be used for decision studies. A series of decision studies was conducted to investigate reproducibility of scores as a function of test length.

For the global ratings a two facet, all random raters nested within stations by persons design [R:(PxS)] was used. A similar pooling of variance components across test administrations was applied as described above. Again, reproducibility of scores as a function of test length and number of examiners was investigated through a series of decision studies. For the final research question, adjusted dependability indices were calculated just as the usual dependability indices, except that squared deviations from the cut-off were used in place of squared deviations from the mean. Implicit in the mastery oriented testing approach is the use of absolute standards in making pass-fail decisions. Traditional approaches to determining cut off points on scoring scales often result in rather arbitrary cut-off values, failing to provide valid evidence for pass-fail decisions. Using expert panellists in standard setting may enhance fairness and defensibility of pass-fail decisions. In our study, we adopted a modified borderline-group method using the global ratings as expert judgements about 'ready-for-practice' level in standard setting. By plotting checklist scores against global rating scores across all stations and examinees, a regression analysis allows estimating checklist scores corresponding to cut-off values on the global rating scale. This 'borderline regression method' provides a relative inexpensive and elegant method for standard setting by a relatively large panel of experts, based on actual

Table 3.1 *Means, standard deviations for checklist scores and global ratings, from sub samples and overall analysis; and correlations between checklist scores and global rating scores*

Test Form	Final checklist score (C)		Overall global rating score (G)		Correlation C-G
	Mean	SD	Mean	SD	
1999 yr 3 N = 26	79.49	4.90	3.21	.42	.650[a]
1999 yr 4 N = 31	75.80	7.12	3.29	.43	.860[a]
2000 yr 3 N = 37	79.56	6.33	2.99	.51	.845[a]
2000 yr 4 N = 15	81.14	5.88	3.25	.41	.770[a]
Overall N = 109	78.69	6.41	3.16	.46	.793[a.b]

[a] All correlations significant at 0.01 level (2-tailed)

[b] Correlations were computed after transforming scores to Z-scores

examinee performance. The borderline regression method differs from the regular borderline method in using all information provided by the global ratings (Muijtjens et al., 2003). In our study, for global rating scores the cut-off value was set at 2 (borderline performance). Using the borderline regression method, the cut-off point for checklist scores was determined by regressing the associated checklist score at borderline performance on the global rating scale, resulting in a cut-off score of 67,57%. For both checklist scores and global ratings we estimated reproducibility of pass-fail decisions as a function of test length (checklists and global ratings) and number of raters (global ratings only).

Results

Comparison of mean checklist scores within each year group did not show any significant differences between the group of students included in the study and the rest of the year group (independent-groups t-test, p > .5, not significant).
Descriptive statistical information is presented in Table 3.1. Table 3.1 presents the means and standard deviations of the score distribution of the final checklist scores and overall global rating scores for the 6 station, 3 hours MCT. Differences between mean scores from test form to test form are minimal and not significant. Correlations between checklist scores and global ratings vary between .65 and .86.

Table 3.2 presents the estimated variance components resulting from generalizability analyses for checklist scores. The variance component for persons (P) indicates how much examinees vary in ability. Pooled across test administrations and years of training, this variance component accounts for about 11% of the total variation in checklist scores. The stations variance component (S), reflecting variance in case (station) difficulty, accounts for about 15% of total variance. As can be seen from Table 3.2, the residual variance component (PS,e) contributes

Table 3.2 *Variance component estimates for the PxS-design and generalizability coefficients for checklist scores as a function of testing time*

Test Form	Variance components*			Generalizability coefficient as a function of testing time		
	P	S	PS,e	3 hrs[a]	5 hrs[a]	10 hrs[a]
1999 yr3 N = 26	5.43	1.60	111.29	0.23	0.33	0.49
1999 yr4 N = 31	25.54	55.37	151.30	0.50	0.63	0.77
2000 yr3 N = 37	20.95	15.67	114.87	0.52	0.65	0.78
2000 yr4 N = 15	21.09	28.32	81.24	0.61	0.72	0.84
Overall N = 109	18.57	25.35	119.75	0.48	0.61	0.76
(percentage of total variance)	(11%)	(15%)	(73%)			

[a] 2 cases per hour

* Variance component estimates for persons (P); stations (S); and residual (PS,e), reflecting variance due to person-by-case interaction (PS) and unidentified sources of error

most to score variance (73%). This large residual variance suggests large examinee by case interaction variance, unidentified sources of measurement error or both.

Table 3.2 also presents estimated generalizability coefficients based on checklist scores as a function of number of cases (testing time). It can be seen that for the 6 station, 3 hours MCT the generalizability coefficient for checklist scores is 0.48, indicating poor generalizability of test results across tasks in the midwifery domain. Very long testing time (more than 10 hours) is needed to achieve near acceptable reliability levels.

Table 3.3 presents variance component estimates and estimated reliability of test scores when using global rating scales as the sole instrument for scoring examinee performance. Comparing information in Table 3.3 with findings on checklist scores, some differences can be noticed. For global ratings, the variance component for persons (P) seems to be somewhat larger than for checklist scores (accounting for 17% of total score variation), indicating better discriminating ability for the global rating scales. Station variance (S) is relatively small (3% of total variance), indicating that differences in station difficulty have only limited influence on global ratings. The variance component reflecting examinee by case interaction (PS) contributes most to variation in scores. This indicates large variation in relative standing of examinees from case to case, reflecting content-specific performance and competence. The residual variance component [R:(PxS),e] is also relatively large, reflecting confounded variation due to rater effects, interactions between raters, persons and stations and unidentified sources of measurement error.

Data in Tables 3.2 and 3.3 furthermore show that the use of global ratings increases reliability of performance scores, with reproducibility coefficients in the moderate range for the 6 station 3hrs MCT. The generalizability coefficient for global ratings is 0.61 for the actual six station MCT (2 raters per case), increasing to 0.84 for a test consisting of 20 stations. From Table 3.3 it can be

Table 3.3 *Variance component estimates for the R:(PxS) design and generalizability coefficients of global rating scores, as a function of testing time and number of raters*

Test Form	Variance components*				Generalizability coefficient as a function of testing time and number of raters					
					3 hrs[a]		5 hrs[a]		10 hrs[a]	
	P	S	PS	R:(PxS),e	1[b]	2[b]	1[b]	2[b]	1[b]	2[b]
1999 yr3 N = 26	0.102	0.023	0.344	0.192	0.53	0.58	0.65	0.70	0.79	0.82
1999 yr4 N = 31	0.087	0.024	0.465	0.231	0.43	0.47	0.56	0.60	0.71	0.75
2000 yr3 N = 37	0.175	0.007	0.389	0.196	0.64	0.68	0.75	0.78	0.86	0.88
2000 yr4 N = 15	0.107	0.024	0.211	0.207	0.61	0.67	0.72	0.77	0.84	0.87
Overall N = 109	0.123	0.018	0.375	0.207	0.56	0.61	0.68	0.72	0.81	0.84
(percentage of total variance)	(17%)	(3%)	(52%)	(29%)						

[a] 2 cases per hour [b] Number of raters per station

* Variance component estimates for persons (P); stations (S); person-by-station interaction (PS); and residual [R:(PxS),e], reflecting variance due to raters (R), interaction between raters, persons and stations and unidentified sources of error

Table 3.4 *Adjusted dependability indices checklist scores as a function of testing time*

Test Form	Cut-off score = 67,57%[a]		
	3 hrs[b]	5 hrs[b]	10 hrs[b]
1999 yr3	0.89	0.93	0.96
1999 yr4	0.71	0.81	0.90
2000 yr3	0.88	0.93	0.96
2000 yr4	0.92	0.95	0.97
Overall	0.85	0.91	0.95

[a] Cut-off score as estimated with borderline regression method in standard setting

[b] 2 cases per hour

seen that increasing the number of cases in the test is far more effective in increasing reproducibility of test results than increasing the number of raters per case.

Tables 3.4 and 3.5 list adjusted dependability coefficients for a mastery oriented score interpretation. Results clearly show that dependability indices with cut-off are larger than the corresponding generalizability indices. For the actual 6 station MCT, the adjusted dependability index for checklist-scores is 0.85 if the cut-off point is set at 67,57% (cut-off point corresponding to global rating level = 2). Similar to findings for generalizability coefficients, adjusted dependability indices for global ratings are larger than for checklist scores. For global ratings, the reproducibility coefficient for decisions increases to 0.95 if the cut-off value is set at 2 (= borderline performance).

Table 3.5 *Adjusted dependability indices global rating scores as a function of testing time and number of raters per case*

Test form	Cut-off score = 2					
	3hrs[a]		5hrs[a]		10 hrs[a]	
	1[b]	2[b]	1[b]	2[b]	1[b]	2[b]
1999 yr3	0.94	0.95	0.96	0.97	0.98	0.98
1999 yr4	0.94	0.95	0.96	0.97	0.98	0.98
2000 yr3	0.92	0.93	0.95	0.96	0.97	0.98
2000 yr4	0.96	0.97	0.97	0.98	0.99	0.99
Overall	0.94	0.95	0.96	0.97	0.98	0.98

[a] 2 cases per hour

[b] Number of raters per case

Discussion and conclusion

The findings from our study are generally consistent with findings in medical literature (Colliver et al., 1989; Van der Vleuten and Swanson 1990; Vu & Barrows 1994; Swanson et al., 1995; Gruppen et al., 1997; Colliver and Swartz, 2000). Depending on score interpretation and the scoring method used, results show variability among measures of consistency as well as among the number of cases required to achieve acceptable reliability levels and time needed to administer that number of cases.

The data on reproducibility of checklist scores indicate low reproducibility of test results across tasks within the domain of midwifery. D-studies show that even with very long tests consisting of more than 20 cases, minimum acceptable levels of reproducibility (0.80) are not achieved. Relative magnitudes of estimated variance components indicate that the case- or content-specificity problem, resulting in score variation from task to task, contributes most to limited reproducibility of test scores. The large examinee-by-case interaction variance for global ratings supports this conclusion. It should be noted that reported generalizability coefficients for checklist scores are in fact an overestimation of test reliability, as checklist scores are based on a consensus score by two raters. Reproducibility coefficients on original independent checklist scores are likely to be even smaller. Our findings on reproducibility of test scores are consistent with findings in medical literature, especially when long stations are used as we did here (Van der Vleuten and Swanson, 1990).

Concerning our first research question, results from our study do not support the assumption that narrowing the task domain of the assessment, in our case from medicine to midwifery, might lead to better reproducibility of scores across tasks by mitigating the content-specificity problem. It remains somewhat counterintuitive that in a domain as small as midwifery, the most important problem in generalisation of performance scores is the problem of task specificity. A possible explanation is the fact that educators constructing competence tests in small professional domains seek to maximise content and context variability across tasks in the test for purposes of construct validity. Maximal content variability across tasks will be reflected in large score variability, especially when task-specific checklists are used. Furthermore it may very well be that, as has been suggested before (Singer et al., 1996), variability in performance across stations is not simply related to content variation, but perhaps to other factors, such as pattern recognition based on irrelevant contextual features of the case. Depending on the case presented to them, students may not be able to screen out irrelevant information. Even if they possess adequate prior knowledge, they are likely to be distracted by irrelevant cues and formulate unnecessary goals in problem solving. Contextual variance may therefore deserve special consideration when assessing competence of students and starting professionals (beginning experts), for whom both relevant and irrelevant contextual features influence approaches to clinical problem solving.

Our findings on the use of global rating scales seem to confirm findings in medical literature with respect to the use of checklists and global ratings in performance based assessments. Based on global ratings alone, test scores show quite reasonable reproducibility given the limited number of cases in the test. Findings indicate that, although the influence of the task-specificity phenomenon on score reproducibility remains considerable, global ratings lead to better reproducibility of overall test scores. Global ratings seem to be more effective in differentiating examinees ability, thereby reducing test length requirements to some extent. Correlation studies between checklist scores and global rating scores provide some evidence for the validity of global ratings in evaluating competence, for both checklists and global rating scales were designed to measure the same thing, namely the ability of the student to use the appropriate knowledge, skills and attitude in handling tasks in the professional domain. However, interpreting correlations between measures is difficult without further research into the validity of the measurement constructs. In addition, our results are confounded by the fact that all raters were intensively trained in scoring performance on the basis of checklist scoring protocols, and both checklists and global ratings were used simultaneously in scoring student performance. For that reason, it is likely that checklist scoring protocols greatly influenced global rating scores representing expert judgements of competence. Informal feedback from examiners (raters), however, confirmed our assumption that checklists have limited validity in assessing competence. Examiners stated that global rating scales also allowed for capturing elements of professional competence that did not appear on the checklist, e.g. coherence in information gathering and efficiency in problem solving. Compared to checklists, which are highly content-specific, the global ratings seem to represent a broad and more stable (task-independent) range of skills necessary for good professional practice.

From our findings it may be concluded that the use of global ratings seems to be a promising alternative to checklists when assessing competence on performance based examinations, especially considering the fact that the use of global ratings would greatly reduce costs in test design and test administration. However, development of formats for accurate feedback to students as well as reliance on global ratings in high-stakes examinations will require more research into validity of global ratings and ways of optimising construction of global rating scales.

From our study it becomes clear that test length requirements may be reduced not only by using global rating scales, but also by adopting a mastery-oriented perspective for score interpretation. Findings from our study show that short tests can yield high consistency levels.

The adjusted dependability index, however, must be interpreted in relation to the chosen cut-off and its distance to the mean. A major problem in interpretation of adjusted dependability indexes is the determination of the appropriate cut-off score. Combining expert ratings of competence level with more detailed scoring of observed performance as described in this article, provides a relative inexpensive and elegant method for standard setting by large panels of experts.

Conclusively, results show that reproducibility of test scores on performance-based tests in midwifery is problematic, although reproducibility can be ameliorated by using global rating scales or adopting a mastery-oriented score interpretation, or both. Generalizability findings indicate that even for small domains, the problem of case-specificity remains a major impediment to implementing performance-based examinations in high stakes tests. Both for analytical and global scoring methods very long test lengths are required for scores to be sufficiently reliable for meaningful interpretation. Findings furthermore suggest that use of global rating scales and professional expert judgements in professional competence evaluation is to be preferred to task specific checklists for reasons of reproducibility (test length requirements) and efficiency in test development and administration. Adoption of a mastery-oriented test perspective, as is most appropriate in licensure or certification procedures, may reduce testing time requirements even further. However, it is also clear that more research into validation of global ratings and defensible standard setting is still needed.

References

Cohen, D.S., Colliver, J.A., Robbs, R.S. & Swartz, M.H. (1997). A large-scale study of the reliabilities of checklist scores and ratings of interpersonal and communication skills evaluated on a standardized-patient examination. *Advances in Health Sciences Education, 1*, 209-213.

Colliver, J.A., Verhulst, S.J., Williams, R.G. & Norcini, J.J. (1989). Reliability of performance on standardized patient cases: A comparison of consistency measures based on generalizability theory. *Teaching and Learning in Medicine, 1*(1), 31-37.

Colliver, J.A. & Swartz, M.H. (2000). Reliability and validity issues in standardized patient assessment. In L.H. Distlehorst, G.L. Dunnington & J.R. Folse (Eds.), *Teaching and learning in medical and surgical education: lessons learned for the 21st century* (pp. 229-236). Mahwah: Lawrence Erlbaum Associates.

Crébas, A. (1989). *Beroepsomschrijving Verloskundigen. Beroepsomschrijving opgesteld in opdracht van de Nederlandse Organisatie van Verloskundigen* (Description of the duties of a midwife, written by order of the Netherlands Organisation of Midwives). Bilthoven.

Cunnington, J.F.W., Neville, A.J. & Norman, G.R. (1997). The risk of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education, 1*, 227-233.

Eraut, M. (1994). *Developing professional knowledge and competence.* London: Falmer Press.

Govaerts, M.J.B., Schuwirth, L.W.T., Pin, A.K., Clement, M.E.J. & Vleuten van der, C.P.M. (2001). Objective assessment is needed to ensure competence. *British Journal of Midwifery, 9*(3), 156-161.

Gruppen. L.D., Davis, W.K., Fitzgerald, J.T. & McQuillan, M.A. (1997). Reliability, number of stations, and examination length in an objective structured clinical examination. In A.J.J.A. Scherpbier, C.P.M. van der Vleuten, J.J. Rethans & A.W.F. van der Steeg (Eds.), *Advances in Medical Education* (pp. 441-442). Dordrecht: Kluwer Academic.

Hodges, B., Regehr, G., McNaughton, N., Tiberius, R. & Hanson, M. (1999). OSCE checklists do not capture increasing levels of expertise. *Academic. Medicine, 74*, 1129-1134.

Kane, M.T. (1992). The assessment of professional competence. *Evaluation & the Health Professions, 15*(2), 163-182.

Kane, M.T. (1994). Validating interpretative arguments for licensure and certification examinations. *Evaluation & the Health Professions*, *17*(2), 133-159.

Kane, M., Crooks, T. & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice,* Summer 1999, 5-17.

Muijtjens, A.M.M., Kramer, A.W.M., Kaufman, D.M. & Van der Vleuten, C.P.M. (2003). Using resampling to estimate the precision of an empirical standard-setting method. *Applied Measurement in Education, 16*(3), 245-256.

Newble, D. (1988). Eight years' experience with a structured clinical examination. *Medical Education, 22*, 200-4.

Newble, D. & Jaeger, K. (1983). The effects of assessments and examinations on the learning of medical students. *Medical Education, 17*, 165-71.

Norcini, J.J. & Swanson, D.B. (1989). Factors Influencing testing time requirements using written simulations. *Teaching and Learning in Medicine, 1*(2), 85-91.

Regehr, G., MacRae, H., Reznick, R.K. & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine, 73*(9), 993-997.

Reznick, R.K. & Rajaratanam, K. (2000). Performance-based assessment. In L.H. Distlehorst, G.L. Dunnington & J.R. Folse (Eds.), *Teaching and learning in medical and surgical education: lessons learned for the 21st century* (pp. 237-244). Mahwah: Lawrence Erlbaum Associates.

Rothman, A.I., Blackmore, D., Dauphinee, W.D. & Reznick, R. (1997). The use of global ratings in OSCE station scores. *Advances in Health Sciences Education, 1*, 215-219.

Singer, P.A., Robb, A., Cohen, R., Norman, G. & Turnbull, J. (1996). Performance-based assessment of clinical ethics using an objective structured clinical examination. *Academic Medicine, 71*(5), 495-498.

Swanson, D.B. & Norcini, J.J. (1989). Factors influencing reproducibility of tests using standardized patients. *Teaching and Learning in Medicine, 1*(3), 158-166.

Swanson, D.B., Norcini, J.J. & Grosso, L.J. (1987). Assessment of clinical competence: Written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12*(3), 220-246.

Swanson, D.B., Norman, G.R. & Linn, R.L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher, 24*(5), 5-11, 35.

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

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

Vu, N.V. & Barrows, H.S. (1994). Use of standardized patients in clinical assessments: Recent developments and measurement findings. *Educational Researcher, 23*(3), 23-30.

CHAPTER 4

The use of observational diaries in in-training evaluation: student perceptions

Published in Advances in Health Sciences Education 2005,
Vol 10, No 3: 171-188

Marjan J.B. Govaerts
Cees P.M. van der Vleuten
Lambert W.T. Schuwirth
Arno M.M. Muijtjens

Abstract

Introduction • In health science education clinical clerkships serve the twofold purpose of guiding student learning and assessment of performance. Evidently, both formative and summative assessment procedures are needed in clerkship assessment. In-training evaluation (ITE) has the potential to serve both assessment functions. Implementation of effective ITE, however, has been shown to be problematic, partly because integration of assessment functions may have negative consequences for teaching and learning. This study investigates student perceptions of the impact of an integrated assessment approach, seeking to refine criteria for effective ITE.

Method • In the curriculum of Maastricht Midwifery School (MMS)[2], clerkship assessment is based on ITE serving both assessment functions. The ITE model is based on principles of extensive work sampling, and frequent documentation of performance. A focus group technique was used to explore student perceptions on the impact of the ITE approach on student learning and supervisor teaching behaviour, and on the usefulness of information for decision making.

Results • Results indicate that the assessment approach is effective in guidance of student learning. Furthermore, students consider the frequent performance documentation essential in clerkship grading. Acceptance and effectivity of ITE requires a learning environment which is safe and respectful. Transparency of assessment processes is the key to success. Suggestions for improvement focus on variation in evaluation formats, improvement of feedback (narrative, complete) and student involvement in assessment.

Conclusion • ITE can fulfil both its formative and summative purposes when some crucial conditions are taken into account. Careful training of both supervisors and students in the use of ITE for student learning and performance measurement is essential.

[2] Maastricht Midwifery School is now known as the Faculty of Midwifery Education and Studies, Zuyd University (Academie Verloskunde Maastricht)

Introduction

In health sciences education, most curricula are characterized by clinical attachments (clerkships or rotations), serving a twofold purpose. First, they aim to provide a learning environment that motivates and guides student learning towards a standard of professional competence. Second, these educational settings provide the best conditions for assessment on the DOES level (Miller, 1990), ensuring that upon completion of the training period the students have reached a 'fit-for-practice' level of competence. In this respect, the importance of both formative assessment (to guide students' learning) as well as summative assessment procedures (for purposes of accountability) is self-evident. In-training evaluation (ITE), defined as the ongoing assessment, documentation and feedback on student performance, is considered to be an assessment strategy which has the potential to serve both formative and summative assessment functions in a real life practice setting (Feletti et al., 1994; Hays and Wellard, 1998; Spike et al., 2000). Assessment of performance by medical educators in the setting of day- to-day practice is agreed to be an invaluable tool in the measurement of clinical competence. Furthermore, feedback on-the-job by professional supervisors is crucial in competence development. However, main weaknesses in current ITE programs are the infrequency of direct observation and documentation of student performance, limiting their value in guidance of student learning, and resulting in questionable reliability of evaluations.

Recently, Turnbull and Van Barneveld (2002) proposed basic requirements for improvement of in-training evaluation. The main characteristics of their ideal ITE-program can be summarized as follows:

- Continuous performance assessment through systematic observation, standardized evaluation and objective documentation of student performance in real clinical settings. An increase in the range and number of situations observed is a primary goal. Whenever possible, different raters should be included in the assessment strategy;
- Assessments that cover all relevant areas of competence -as specified in clerkship objectives and expectations for performance;
- Acceptability of assessment, requiring the approach to be flexible, feasible and viewed as an important tool for decision making, identifying learning needs and providing feedback.

In this way, ITE meets essential preconditions for both summative and formative evaluation. Clerkship assessment will be based on extensive work sampling and performance assessment in ecologically valid situations, which are fundamental requirements for summative decision making. Furthermore, within this approach, feedback can be expected to be effective (Ende, 1983; Rolfe and McPherson, 1995). Feedback can be linked to explicit goals, giving concrete information about students' progress in professional development. Also, feedback will be complete, based on first-hand data and provided frequently, while remedial action is readily available.

From a theoretical perspective, this ITE approach seems a promising alternative to evaluate student performance in clinical clerkships. However, an important question concerns the adequacy of the assessment approach for the proposed purposes: can ITE indeed meet its goals in educational practice? For example, feasibility issues often limit the usefulness of ITE-approaches. Supervisors do not often take the time to observe students or document performance measurements. Furthermore, it is well known that assessment methods and the use of assessment results affect student and teacher behaviour. For instance, research on multiple station examinations has shown that students adapt their learning behaviour and performance depending on perceptions of how their performance will be scored (Van Luijk et al., 1990; McIlroy et al., 2002). Also, several authors have suggested that the combination of formative and summative assessment functions may interfere with the relationship between supervisor and student (Hays and Wellard, 1998; Sommers et al., 1994; Toohey et al., 1996). Students may be unwilling to admit to anxieties or shortcomings in their performance in case it influences their future assessments. Supervisors, on their part, will be reluctant to make any negative comments about a student or report their performance as unsatisfactory. In this way the combination of formative and summative assessment functions may have a negative impact both on learning outcomes and usefulness of performance documentations for summative decision making. Therefore, it is important to seek evidence about how the assessment actually works within the system it is used, studying intended and unintended consequences of the assessment method. Investigation of interpretations and perceptions of assessment practices must play a major role in this validation process.

The present article describes student perceptions of the educational impact of an in-training evaluation model integrating formative and summative assessment, based on the principles of extensive work sampling and continuous documentation of student performance. Focus group interviews were conducted as an initial, exploratory research into the consequences of the integrated ITE in midwifery clerkships. The study focuses on the effects of the assessment method on student learning; the perceived significance of the assessment instrument in decision making and clerkship grading, and effects of the integrated assessment on student and supervisor behaviour.

Method

Context of the study

This study was conducted in the Netherlands, within the setting of midwifery education. The curriculum of Maastricht Midwifery School (MMS) is a 4-year training program including attachments in primary health care (ambulatory setting) as well as in hospital settings. During the “primary health care” clerkships students are attached to a midwifery practice, one or several midwifes acting both as teacher (supervisor) and evaluator of performance. In in-patient settings students are supervised by various health care staff members (obstetricians, residents, midwives as well as nurses). After finishing midwifery school, a midwife is authorised to independently practice midwifery in primary health care.

Assessment in MMS midwifery clerkships

Since 1997, MMS has been using a model for clerkship assessment in which frequent formative assessments provide evidence for summative performance evaluation, thus aiming to meet the dual function of the clerkship.

Essential characteristics of this model reflect basic ITE requirements:

- Clear setting of learning goals and performance standards at the beginning of each clerkship;
- Frequent (daily) direct observation of the student when participating in midwifery care, followed by;
- Frequent (daily) evaluation and documentation of student performance, registration of feedback;
- *All* supervisors who work with the student in day-to-day practice are required to observe and document student performance;
- Summative decision making on the basis of evidence collected throughout the training period.

Students and supervisors are provided with instruments to support realization of assessment requirements, including study guides and observational diaries. The study guides outline the educational goals and performance standards for each of the clerkships. In addition, they describe assessment procedures and they require students and supervisors to define mutual expectations and teaching/learning agendas at the beginning of the clerkship. Standardization of evaluation is attained by the use of "direct observation forms" in the observational diary, each form consisting of a checklist of performance items covering several areas of midwifery competence (for an example see Figure 4.1). Judgments on performance are made on a three-point scale (unsatisfactory, borderline, satisfactory), against pre-set criteria (descriptions of performance level to be achieved at the end of the clerkship). Each form further provides ample space for registration of additional comments and feedback in writing. It is the responsibility of the student to ensure daily completion of the assessment forms. Both students and supervisors receive instruction and training in the use of the study guide and observational diaries.

Primary purpose of this instrument is to structure student learning throughout the clerkship. Also, the mandatory daily evaluation of student performance and registration of feedback provides opportunities for student-preceptor interaction and reflection, considered essential for learning in the workplace. Furthermore, evidence collected this way is to be used qualitatively towards a decision about the student's midwifery competence at the end of the clerkship. In performance appraisal, supervisors and the MMS examination board collectively use interpretations of provided documentation on student performance and other information to form a motivated summative decision.

Data collection and analysis

Focus groups were used as a technique to explore students' opinions and experiences with the integrated assessment in midwifery clerkships. Focus groups are a research method for qualitative data gathering through group discussion. The main purpose of this method is to draw upon respondents' attitudes, experiences and reactions which are less likely to be revealed by questionnaire surveys or one-to-one interviews (Morgan, 1997).

Participants were recruited from 4th (final) year student-midwives in the classes of 2001 and 2002 (40 and 29 students respectively). All students were invited to participate in a focus group session. From these classes, 15 students volunteered to participate, from which three focus groups were created. Eventually, three students were not able to join their focus group session

Student name: ……………..
Supervisor: …………….. Clerkship period: ……………...

	Date:			
	Satisfactory	Not satisfactory	Borderline	Not observed
Management of first stage of labour				
Uterine contractions	0	0	0	0
Assessment and interpretation (progression; abnormalities)				
Foetal heart rate				
Regular and timely auscultation	0	0	0	0
Skill –technique	0	0	0	0
Interpretation	0	0	0	0
Vaginal examination				
Skill – technique	0	0	0	0
Interpretation of findings	0	0	0	0
…………………..	0	0	0	0
…………………..	0	0	0	0
Risk assessment and management first stage	0	0	0	0
Management of second stage of labour				
Foetal heart rate	0	0	0	0
Maternal expulsive efforts	0	0	0	0
Instructions to the woman	0	0	0	0
Amniotomy	0	0	0	0
Analgesia	0	0	0	0
Risk assessment and management second stage	0	0	0	0
Child delivery				
Delivery of the head	0	0	0	0
Delivery of shoulders	0	0	0	0
Nuchal cord(coil); clamping of cord	0	0	0	0
Aspiration nose –mouth	0	0	0	0
APGAR score	0	0	0	0
………………………………	0	0	0	0
………………………………	0	0	0	0
Management of third-fourth stage				
Signs of placental separation/interpretation	0	0	0	0
Delivery of placenta	0	0	0	0
Height of uterine fundus/blood loss/ consistency	0	0	0	0
Examination placenta/membranes/umbilical cord	0	0	0	0
Inspection birth canal / lacerations	0	0	0	0
Risk-assessment and management third/fourth stage	0	0	0	0
Repair of episiotomy				
Instrumentation	0	0	0	0
Knowledge anatomy	0	0	0	0
Technique	0	0	0	0
……………………………	0	0	0	0
……………………………				
Interpersonal skills/communication	0	0	0	0
……………..	0	0	0	0
Documentation	0	0	0	0
Professionalism	0	0	0	0
……………….				
	Initials Supervisor:			

Figure 4.1 Direct observation form natal care (exemplary selection of items)

on the scheduled date, so a total of 12 students (17%) participated in the study, with focus groups of five, three and four students respectively. Students were paid a small compensation for their participation.

Focus group discussions targeted on student perceptions of the effects of observational diaries, regarding the dual function of the assessment model. During the sessions three key topics for discussion were introduced:

- the effects of the observational diary on student learning and behaviour;
- the effects of the observational diary on teaching and supervisor behaviour;
- the usefulness of the observational diary in decision making (grading of the clerkship).

During discussion of these topics, students were also invited to suggest ways to improve the instrument.

The focus groups convened for one session of 1½-2 hours. One of the authors (MG), and a MMS teacher, not involved in clinical teaching, conducted the focus groups. The focus group sessions were tape-recorded with permission from the participants. Participants were assured that their names would remain anonymous. MG introduced the topics for discussion and facilitated group discussion, while the other interviewer acted as an assistant moderator, taking notes and asking for explanations and further comments when appropriate. The tape recordings of the focus group sessions were transcribed.

The transcripts were analysed by the first moderator. First, students' opinions on the clerkship assessment model were categorised according to the key topics in the discussion. Students' ideas were further broken down into subcategories representing different aspects of each topic, as discussed by the students. Subsequently, the information was reordered and summarized in a report. The report, representing the group's discussion, was mailed to the students for approval. All students indicated that the reports accurately represented the discussion.

Results

In the following section students' opinions on the topics discussed in the focus groups will be summarised according to the (sub)categories that were used to analyse the transcripts of the focus group sessions. Quotations that most accurately depict the discussions are provided. Figures indicating student and focus group number are given in parentheses.

Feelings about the effects of the use of the observational diaries were largely consistent between groups, although within each focus group individuals sometimes expressed different opinions or experiences. Where indicated, these differences are noted.

Effects of the observational diary on student learning and behaviour

Guidance of the learning process • All students underline that essential learning takes place during clerkships. Although students have varying perceptions about the usefulness of the observational diary, most students say that criteria in the checklist and listing of specific knowledge, skills and attitudes are helpful in structuring the learning process. It helps students in focussing on essential learning goals. Furthermore, the frequent assessment and feedback regulate the student's learning by giving insight into the way supervisors will judge their performance.

Many students, however, indicate that the detailed checklist for observation and evaluation of performance is useful only in the beginning of the training period, when student learning focuses on application of knowledge and skills, and development of fluency on standard tasks. Students feel that as they are getting more and more experienced, the standardized checklists no longer meet their needs and, with regard to routine tasks, a holistic approach to competence is more appropriate. Furthermore, at the end of the training period, learning goals shift towards skills such as professionalism, organisation of work, efficiency and independent practice, for which no checklists are available. As a consequence, many items in the checklists are perceived as superfluous whereas other, more relevant items are felt to be missing. In this way, the standardized checklists do not reflect individual learning goals and become less useful as a guide for learning.

> *[The observational diary] is like a manual, giving insight into both clerkship objectives and individual learning goals. You are forced to explicate weaknesses in your performance, which automatically leads to new learning goals. (S3,2)*

> *By using the checklists in performance evaluation, the supervisor is forced to assess the same performance items again and again, from the beginning until the end of the clerkship period, even after the student has sufficiently shown that she is competent on some items. Then, performance evaluation becomes meaningless and a waste of time. (S1,3)*

In addition, all students underline that the additional feedback in writing is far more important for their learning process than the checklist-evaluation. It is this feedback that justifies the scores on the checklist, thereby identifying the specific learning goals for the student. Furthermore, the feedback gives insight into the case-specific difficulties of the observed situation. Virtually all students reported that content-specific task information and details about student performance are indispensable for adequate learning (and grading!) during the clerkship. Naturally, standardized checklists cannot provide this relevant information.

> *Personally, I attach far more value to the feedback in writing. The checklist does not reveal the specific issues discussed in this additional feedback. If the feedback in writing was omitted from the observation form, I would not be able to get an adequate impression of my performance. (S2,3)*

Student involvement in the learning process • All students emphasize that self-assessment and reflection on performance is essential for adequate learning. Students regret that the diary does not stimulate them to assess their performance before the supervisor's evaluation, nor provides the opportunity to write down reflective remarks on their performance. To their opinion, incorporation of self-assessment forms in the diary would stimulate students to reflect on the supervisor's feedback and increase involvement in their own learning process. Self-evaluation helps in accepting and using of feedback from the supervisor, and more importantly, comparing evaluation results creates opportunities to discuss learning needs and set up mutually-agreed-upon expectations.

> *In the observational diary, there should be space for reflective comments by the student leading to specific learning goals with every situation. Only then the diary will become an actual learning document. (S3,1)*
>
> *The student must have a chance to fill out a self-assessment form first (….) and compare results with feedback from the supervisor. (….). Then, as a student, you will control your own learning process. Now you will just sit and await the supervisor's judgement. (S1,3)*

Monitoring of the learning process • Students indicate that awareness of progression in relevant competencies and increasing self-confidence are indicative of successful clerkship training. Several students spontaneously mentioned the usefulness of the diary in this respect. They regularly review the diary to check their progression in competence development, by analysing the evaluation results and feedback over a period of time. For these students decreasing frequency of negative feedback indicates increasing professional proficiency. Other students indicated that the diary is not useful in monitoring competence development. Main reasons for the perceived limited use of the diary in guiding and monitoring professional development are the analytic evaluation approach in the checklists and frequency of evaluation (too high). Additional arguments are subjectivity in the evaluation process and incompleteness of feedback on the observation forms.

> *The observational diary only reports evaluations of performance on various occasions, separate performances being judged as adequate, borderline or substandard. As a student you have no idea what these judgements mean in terms of achievement over time and competence development towards end-of-clerkship standards. (S3,2)*

Conflicts of interest and student behaviour • The dual function of the diary, serving as a tool for learning and summative decision making, is considered as non-problematic by most students, even although students report that they feel dependent on the supervisor for their clerkship grading. More than that, some students stress that combining formative and summative assessment is fundamental to meaningful assessment in clerkships. All students however stress that a relationship based on mutual trust and respect is an essential precondition. The communication between student and supervisor has to be open and honest, and the student needs to feel safe in the educational setting, knowing that mistakes are accepted as opportunities to learn. Students feel that these conditions can only be met in situations in which the number of supervisors is limited and continuity in supervision is guaranteed, as is the case in most primary health care settings. In clerkship settings in which many supervisors are involved in performance evaluation, students feel that assessment procedures become less transparent and they may feel less safe. As a consequence, they may withhold information that reveals weaknesses in performance. Students feel that in these settings supervision lacks continuity and assessment procedures become less transparent. Students also indicate that the feeling of dependence occasionally hampers individual learning processes, because students will comply with notions and behaviour of their supervisor even if this is in conflict with what they have learned at school or with their own notions about midwifery practice. Students feel that copying their supervisor's behaviour in this way, may interfere with development of professional competence.

> *The most important condition for clerkship learning is a relationship based on mutual trust, which enables you to feel safe, with opportunities to learn from mistakes, a relationship in which communication is honest and open. A situation in which you can develop your own professional concept of midwifery care…….. (S1,1)*

> *There is no other way to do this [i.e. evaluation of student performance] Whether supervisors can be both teacher and assessor, depends on the way they provide feedback. (S3,2)*

Effects of the observational diary on teaching and supervisor behaviour

Direct observation and feedback • Students report that all supervisors frequently observe them during work and use the diary to register feedback on performance. Especially in the beginning of the clerkship, most students are observed daily, followed by checklist-evaluation. The checklist is considered practical and helpful in structuring performance evaluation and feedback. In the beginning of the clerkship and in non-routine tasks, ample additional feedback is provided with the checklist. However, all students report that supervisors' feedback in writing mainly focuses on weaknesses in student performance (negative feedback), neglecting positive feedback. Compared to negative feedback, positive feedback is given less frequently and does not appear on the form. Apparently, supervisors assume that positive feedback is of less importance for the student in guidance of the learning process. Students however indicate that positive feedback, and explicit documentation of attainment of learning goals, is essential for their learning and self-confidence.

> *Without the structure of the observational diary, performance evaluation would not be as frequent as it is now, and you would easily miss/skip items from the assessment procedure. The diary is an excellent stimulus for evaluation. (S4,3)*

> *To me it is very important that supervisors' feedback is complete and honest, and especially contains positive remarks on performance. Most often feedback is limited to an enumeration of performance items that were substandard (……) Too often supervisors do not realise that positive feedback leads to an increased learning output. (S4,1)*

Use of the diary in teaching • Students report that only few preceptors actively use the diary in their teaching. Students expect preceptors to read the documentation on performance and to integrate feedback and learning goals in the planning of learning experiences or the way they coach/supervise students during work. Especially in settings in which students are supervised by a large number of teachers, students may therefore experience lack of continuity and structure in their learning processes. In these settings, adequate use of the assessment instrument for student learning requires intensive consultation between supervisors, which is not always possible.

> *Actually, a supervisor should read over the feedback in the diary before the next observation and performance assessment. Then she should discuss the way in which she intends to support the student and give some suggestions for performance improvement. But as a student you must always remind supervisors to do that. I never experienced a supervisor asking for my diary to check how things are/I'm getting on. (S3,2)*

Conflicts of interest and supervisor behaviour • Students note that discrepancies between judgements and ratings may occur. Especially in their *documentation* of student performance, supervisors appear to be guided by impressions of the student's competence level based on earlier observations or evaluation results in the diary. Occasionally, this may result in under- or overrating of student performance in the observational diary. For some supervisors, the fact that information in the diary is regarded as evidence for clerkship grading is a decisive factor in the way evaluation of performance is documented. On some occasions, substandard performance may then be scored as 'sufficient' in checklist-evaluation.

> *When evaluating my performance, a supervisor once told me that she thought it was substandard in the case concerned, but that she would mark it as satisfactory on the form. She did not want me to fail the clerkship because of unsatisfactory marks in the diary, because, to her opinion, my overall performance was okay. (S1,2)*

Usefulness in summative decision making and grading

About the usefulness of the observational diary for summative decisions during clerkships, students' opinions are ambivalent. Students feel that for summative purposes the observation forms should provide all the necessary evidence regarding their performance during the clerkship. Students, however point out that, for several reasons, the observational diary currently does not fulfil its purpose in summative assessment. Students argue that documentation of student performance is often inaccurate and incomplete, because of the negligence of essential skills in the checklist and the focus on negative feedback, as described before. Students also state that the use of the checklist and three-point rating scale leads to a simplistic and undifferentiated picture of professional competence. According to students, judicious interpretation of checklist scores is very difficult, especially when additional relevant information about the case and student behaviour is lacking.

Furthermore, students feel that evaluation still remains subjective, despite the availability of training, instruction and criteria in the study guides and observational diaries. Supervisors may have highly personal ideas about professional quality, influencing interpretation of performance criteria and standards. Students also note that there are different conceptions about performance scoring: some supervisors adjust their scoring to the momentary stage of development of the student, whereas others use the end-of-clerkship standards as the directive criterion. Students feel that clerkship grading would become more transparent and fair if the diary would compel supervisors to regularly write down a holistic and substantiated judgement on professional competence of the student.

> *It is called an observational diary, but supervisors are forced to give normative statements. They have to mark performance as satisfactory or unsatisfactory. These are value judgements, not observations. …… Every supervisor expects different things from you as a student. (S1,2)*
>
> *(….) overall impressions of competence development cannot be deduced from the separate evaluation forms, especially if there are many supervisors. Quality of performance will vary, depending on the problems presented. How to interpret evaluation results, then? (S2,3)*

Discussion and conclusion

The results from our study seem to indicate that the integrated in-training-assessment model as described for midwifery clerkships is feasible and flexible in use. It is accepted as an important tool for clerkship learning as well as clerkship grading, provided that some crucial conditions are taken into account. A significant finding from our study is the fact that acceptance –and therefore effectiveness of ITE- requires a learning environment in which students feel safe and respected. This is in line with earlier findings on excellence in clinical teaching and supervision (Irby, 1995; Kilminster and Jolly, 2000). Results from our study seem to indicate that transparency of the assessment process is the key to success. Honesty/openness and clarity in communication are essential preconditions for motivation of student learning as well as acceptance of ITE in summative decision making. Especially in settings in which many supervisors are involved in student evaluation, effective ITE therefore places great demands on teamwork to ensure continuity in the coaching of student learning and performance assessment.

Regarding its formative function, our study suggests that the MMS assessment method is effective in the structuring and guidance of student learning in midwifery clerkships, through the high frequency evaluations and using the observational diaries. The MMS assessment method heavily relies on the use of detailed checklists. Although the use of detailed instruments for summative assessments has been challenged by many researchers (e.g. Van der Vleuten et al., 1991; Norman et al., 1991; Regehr et al., 1998; Hodges et al., 1999; Govaerts et al., 2002), checklists are considered superior in formative assessment (Gray, 1996; Cushing, 2002; Imseis and Galvin, 2004). Detailed checklists are considered to ensure specific and concrete feedback, focussing on essential skills and expected behaviours. However, our findings may shed a different light on the usefulness of standardized checklists for student learning. Motivation of student learning requires assessment procedures to be continuously perceived as meaningful and reflecting individual learning goals. As a consequence, the usefulness of standardized checklists listing detailed performance items is limited to the beginning of a training period when student learning focuses on acquisition of routine procedures. As students get more proficient, achieving competence in tasks concerned, holistic judgements are more appropriate. These findings are consistent with earlier research on expertise development in medicine, showing that as learners make progress through different levels of schooling, their performance develops from mastery of individual skill towards holistic competencies (e.g. Eraut, 1994). However, to maintain beneficial effects on student learning, the holistic assessments have to be completed with listings of learning goals related to individual competence development. In this way, student learning will get focused on an individual learning agenda, keeping up motivation and active involvement in learning.

Results from our study underline feedback as the key factor in student learning. Improvement of the formative function of in-training evaluation requires a focus on descriptive (narrative) feedback in evaluation results. Student learning is especially directed by the feedback in additional comments with the checklist. This feedback is more effective than evaluative checklist marks, as it particularly takes into account individual learning goals and task-specific information. In addition, effectiveness of feedback increases when it contains concrete and specific positive information in addition to negative feedback. Positive feedback in documentation of performance is essential for motivation of students, acquisition of self confidence and monitoring of the learning process by focusing attention to mastering of tasks. These suggestions for improvement of ITE regarding its formative function are in line with findings on effectiveness of feedback (Brinko, 1993). Finally, for students assessment becomes more meaningful, resulting in more active learning, when they are engaged in assessment procedures. Self-assessment is fundamental to new concepts of professional competence which include reflective practice and self-directed learning (Stewart et al., 2000; Hays et al., 2002). The use of self-assessment in competence assessment, however, is not undisputed. Studies reporting on accuracy and reliability of self-assessments consistently seem to indicate that professionals' and students' ability to self-assess is questionable (Falchikov and Boud, 1989; Gordon, 1991). The process of self-assessment is complicated and it is evident that self-assessment skills are no natural gift nor achieved by informal training (Eva et al., 2004). However, there is encouraging evidence that explicit training strategies can result in improvement of self-analysis of clinical performance, especially under conditions where learners are involved in collection and interpretation of performance data and if these interpretations are reconciled with supervisors' judgments (Gordon, 1992). Incorporation of self-evaluation in in-training evaluation may provide excellent opportunities for students to analyse their performance by comparing self-reports with expert scores and feedback. Used in this way, it may contribute to high quality learning, better performance and development of essential professional competencies (Topping, 2003).

With regard to clerkship grading, students indicate that summative decision making must rely on information in the observational diary, ITE-results providing essential evidence for competence in real life practice. However, students seem to perceive the MMS in-training evaluation method as 'unfair', lacking in clarity. Suggestions for improvement of the summative function of ITE focus on substantiation of judgments and transparency in assessment. Remarkably, students' comments concern some fundamental preconditions for summative assessment i.e. the multitude of occasions and observers. The MMS assessment model (very frequent evaluation, detailed checklists, non-differentiating rating scale) leads to fragmentation of competence, failing to reflect both holistic judgements and progress in learning. Similar to performance of experienced professionals, quality of student performance can be expected to show (large) variations over time, reflecting both high-level learning and content-specificity of tasks

(Handfield-Jones et al., 2002). Assessment practices should acknowledge the complexity of learning in practice. Therefore, normative statements about student's performance should be given only at intervals. These statements should reflect progress toward end-of-clerkships goals and must be substantiated by collected data in the diary.

Despite rater training and descriptions of performance criteria, great variations continue to exist among supervisors with regard to ratings of student performance. Personal ideas about professional quality and beliefs about student motivation in clerkship learning underlie the way the instrument is used in performance evaluation. These findings reflect well-known phenomena in research on performance appraisal, showing that organizational and motivational factors influence rater behaviour (Cleveland and Murphy, 1992). The need to intensify training of supervisors to evaluate students according to MMS guidelines seems obvious. However, from the student perspective a shift from normative checklist evaluation towards more qualitative, descriptive evaluations is more important and effective in remediation of 'unfairness'. Descriptions of student performance explicitly related to learning objectives and supervisor's interpretations of standards throughout the clerkship period, will contribute to openness and clarity of assessment. In this way, documentation of student performance will provide information from which decisions in the assessment process can be traced and confirmed. Other effective ways to make assessment more transparent to students is to engage them in the assessment process through discussion of assessment criteria and performance standards, and through self assessment practices (Gielen et al., 2003). This will lead to shared understanding of performance and performance standards, contributing to meaningfulness and credibility of in-training evaluation.

Results from our study indicate that students consider the combination of formative and summative assessment functions as inherent in clerkship evaluation and acceptable provided conditions of transparency in assessment are met. Adaptive behaviour in response to the summative use of feedback does occur, but is infrequent. Incorporation of assessment in daily practice, as in the MMS approach, might be a crucial facilitating factor in acceptance of the integrated assessment. Performance judgement and reflection on professional behaviour become part of daily routine, reducing stress inherent to summative assessments, each assessment being followed by many opportunities for performance improvement.

Data collection in our study took place through focus group discussions. Limitations of our study relate to it being a small scale exploratory study. This may limit generalizability of our findings to other settings. Also, our study sampling was based on the use of volunteers, which may have induced bias into the findings. The primary purpose of our study was to gain in-depth understanding and to learn from student perceptions of the MMS assessment method. Focus groups are a highly effective method for gaining detailed information, not otherwise obtained using other methods. Three student groups from two different year groups identified very similar strengths and weaknesses of the ITE-approach. Therefore, with the limitations of the study acknowledged, the authors feel that the present study offers valuable information for designing ITE and suggestions for further improvement of this assessment method.

Results from our study seem to indicate that, to optimise acceptance and effectiveness, approaches to in-training evaluation should:

- Be part of daily practice;
- Balance standardisation (in attempts to increase reliability) with flexibility (to attune individuality in learning processes). ITE should incorporate a range of evaluation formats, standardization of evaluation becoming less important with progression in expertise;
- Promote feedback to be complete, i.e. ensure documentation of weaknesses and strengths in student performance;
- Focus on qualitative, descriptive feedback as a key factor in student learning and transparency of ITE for summative purposes;
- Incorporate self-assessment practices, promoting both reflective practice and discussions about performance interpretations. In addition to complete and descriptive feedback, interaction about performance evaluation will contribute to a shared understanding and agreement for all parties involved in the assessment process;
- Require normative statements of student performance to be given at intervals, to be substantiated by information from daily performance evaluations.

Even more important than characteristics of the assessment instrument is the educational setting in which ITE is to be used. Well informed and well trained supervisors and students are basic requirements for effective integration of formative and summative functions in ITE.

Recommendations for further research include confirmation of our findings in other educational settings. Other research questions should address effects of refinements as described above on feasibility and acceptance of ITE; and strategies to improve training of supervisors and students in the use of in-training evaluation methods. Finally, the question may be raised what the impact can be of including raters with different backgrounds (e.g. patients or peers) on the effectiveness of ITE.

References

Brinko, K.T. (1993). The practice of giving feedback to improve teaching. *Journal of Higher Education, 64*(5), 574-593.

Cleveland, J.N. & Murphy, K.R. (1992). Analyzing performance appraisal as goal-directed behaviour. In G. Ferris & K. Rowland (Eds.), *Research in personnel and human resources management* (Vol. 10, pp. 121-185). Greenwich, CT: JAI Press.

Cushing, A. (2002). Assessment of non-cognitive factors. In: G.R. Norman, C.P.M. van der Vleuten & D.I. Newble (Eds.), *International Handbook of Research in Medical Education* (pp. 711-755). Dordrecht: Kluwer Academic Publishers.

Ende, J. (1983). Feedback in clinical medical education. *JAMA 250*, 777-781.

Eraut, M. (1994). *Developing Professional Knowledge and Competence*. London: Falmer Press.

Eva, K.W., Cunnington, J.P.W., Reiter, H.I., Keane, D.R. & Norman, G.R. (2004). How can I know what I don't know? Poor self assessment in a well-defined domain. *Advances in Health Sciences Education, 9*, 211-224.

Falchikov, N. & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*, 395-430.

Feletti, G., Cameron, D., Dawson-Saunders, B., des Groselliers, J-P., Farmer, E. & McAvoy, P. (1994). In-training assessment. In D. Newble, B. Jolly & R. Wakeford (Eds.), *The Certification and Recertification of Doctors* (pp. 151-166). Cambridge: Cambridge University Press.

Gielen, S., Dochy, F. & Dierick, S. (2003). Evaluating the consequential validity of new modes of assessment: The influence of assessment on learning, including pre-, post-, and true assessment effects. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising New Modes of Assessment: In Search of Qualities and Standards* (pp.37-54). Dordrecht: Kluwer Academic Publishers.

Gordon, M.J. (1991). A review of the validity and accuracy of self-assessments in health professions training. *Academic Medicine, 66*, 762-769.

Gordon, M.J. (1992). Self-assessment programs and their implications for health professions training. *Academic Medicine, 67*, 672-679.

Govaerts, M.J.B., Vleuten van der C.P.M. & Schuwirth, L.W.T. (2002). Optimising the reproducibility of a performance-based test in midwifery. *Advances in Health Sciences Education, 7*, 133-145.

Gray, J.D. Global rating scales in residency education. (1996). *Academic Medicine, 21*(1suppl), S55-63.

Handfield-Jones, R.S., Mann, K.V., Challis, M.E., Hobma, S.O., Klass, D.J., McManus, I.C., Paget, N.S. et al. (2002). Linking assessment to learning: A new route to quality assurance in medical practice. *Medical Education*, *36*, 949-958.

Hays, R. & Wellard, R. (1998). In-training assessment in postgraduate training for general practice. *Medical Education, 32*, 507-513.

Hays, R.B., Jolly, B.C., Caldon, L.J.M., McCrorie, P., McAvoy, P.A., McManus, I.C. et al. (2002). Is insight important? Measuring capacity to change performance. *Medical Education, 36*, 965-971.

Hodges, B., Regehr, G., McNaughton, N., Tiberius, R. & Hanson, M. (1999). OSCE checklists do not capture increasing levels of expertise. *Academic Medicine, 74*, 1129-1134.

Imseis, H.M. & Galvin, S.L. (2004). Faculty and resident preference for two different forms of lecture evaluation. *American Journal of Obstetrics and Gynecology, 191*(5), 1815-1821.

Irby, D.M. (1995). Teaching and learning in ambulatory care settings: A thematic review of the literature. *Academic Medicine, 70*(10), 898-931.

Kilminster, S.M. & Jolly, B.C. (2000). Effective supervision in clinical practice settings: A literature review. *Medical Education, 34*, 827-840.

McIlroy, J.H., Hodges, B., McNaughton, N. & Regehr, G. (2002). The effect of candidates' perceptions of the evaluation method on reliability of checklist and global rating scores in an Objective Structured Clinical Examination. *Academic Medicine*, *77*, 725-728.

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

Morgan, D. (1997). *Focus Groups as Qualitative Research*. Thousand Oaks etc: Sage Publications.

Norman, G.R., Vleuten van der, C.P.M. & Graaf de, E. (1991). Pitfalls in the pursuit of objectivity: Issues of validity, efficiency, and acceptability. *Medical Education, 25*, 119-126.

Regehr, G., MacRae, H., Reznick, R.K. & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine, 73*(9), 993-997.

Rolfe, I. & McPherson, J. (1995). Formative assessment: How am I doing? *The Lancet, 345*, 837-839.

Rothman, A.I., Blackmore, D., Dauphinee, W.D. & Reznick, R. (1997). The use of global ratings in OSCE station scores. *Advances in Health Sciences Education, 1*, 215-219.

Sommers, P.S., Muller, J.H., Saba, G.W., Draisin, J.A. & Shore, W.B. (1994). Reflections-on-action: Medical students' accounts of their implicit beliefs and strategies in the context of one-to-one clinical teaching. *Academic Medicine, 69*(10), S84-6.

Spike, N., Alexander, H., Elliott, S., Hazlett, C., Kilminster, S., Prideaux, D. & Roberts, T. (2000). In-training assessment – its potential in enhancing clinical teaching. *Medical Education, 34*, 858-61.

Stewart, J., O'Halloran, C., Barton, J.R., Singleton, S.J., Harrigan, P. & Spencer, J. (2000). Clarifying the concepts of confidence and competence to produce appropriate self-evaluation measurement scales. *Medical Education, 34*, 903-909.

Toohey, S., Ryan, G. & Hughes, C. (1996). Assessing the practicum. *Assessment and Evaluation in Higher Education, 21*(3), 215-227.

Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In: M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising New Modes of Assessment: In Search of Qualities and Standards* (pp. 55-89). Dordrecht: Kluwer Academic Publishers.

Turnbull, J. & Barneveld van, C. (2002). Assessment of clinical performance: In-training evaluation. In: G.R. Norman, C.P.M. van der Vleuten & D.I. Newble (Eds.), *International Handbook of Research in Medical Education* (pp.793-810). Dordrecht: Kluwer Academic Publishers.

Van Luijk, S.J., Vleuten van der, C.P.M. & Schelven, R.M. (1990). The relation between content and psychometric characteristics in performance-based testing. In W. Bender, R.J. Hiemstra, Scherpbier. A.J.J.A. & Zwierstra R. P. (Eds.), *Teaching and Assessing Clinical Competence* (pp.202-207). Groningen: Boekwerk Publications.

Vleuten van der, C.P.M. Norman, G.R. & Graaf de, E. (1991). Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education, 25*, 110-118.

Vleuten van der, C.P.M., Scherpbier, A.J.J.A., Dolmans, D.H.J.M., Schuwirth, L.W.T., Verwijnen, G.M. & Wolfhagen, H.A.P. (2000). Clerkship assessment assessed. *Medical Teacher, 22*(6), 592-600.

CHAPTER 5

Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment

Published in Advances in Health Sciences Education 2007,

Vol 12, No 2: 239-260

Marjan J.B. Govaerts
Cees P.M. van der Vleuten
Lambert W.T. Schuwirth
Arno M.M. Muijtjens

Abstract

Context • In-Training Assessment (ITA), defined as multiple assessments of performance in the setting of day-to-day practice, is an invaluable tool in assessment programmes which aim to assess professional competence in a comprehensive and valid way. Research on clinical performance ratings, however, consistently shows weaknesses concerning accuracy, reliability and validity. Attempts to improve the psychometric characteristics of ITA focusing on standardization and objectivity of measurement thus far result in limited improvement of ITA-practices.

Purpose • The aim of the paper is to demonstrate that the psychometric framework may limit more meaningful educational approaches to performance assessment, because it does not take into account key issues in the mechanics of the assessment process. Based on insights from other disciplines, we propose an approach to ITA that takes a constructivist, social-psychological perspective and integrates elements of theories of cognition, motivation and decision making. A central assumption in the proposed framework is that performance assessment is a judgment and decision making process, in which rating outcomes are influenced by interactions between individuals and the social context in which assessment occurs.

Discussion • The issues raised in the article and the proposed assessment framework bring forward a number of implications for current performance assessment practice. It is argued that focusing on the context of performance assessment may be more effective in improving ITA practices than focusing strictly on raters and rating instruments. Furthermore, the constructivist approach towards assessment has important implications for assessment procedures as well as the evaluation of assessment quality. Finally, it is argued that further research into performance assessment should contribute towards a better understanding of the factors that influence rating outcomes, such as rater motivation, assessment procedures and other contextual variables.

Introduction

In medical education, the growing interest in direct performance assessment in recent decades has prompted the development of a wide range of 'authentic' assessment methods. Until recently, the 'authenticity movement' focused primarily on performance-based assessment that relied on simulations of complex professional practice, such as OSCEs, standardised patient techniques and computerised patient management simulations (Van der Vleuten, 1996; Reznick and Rajaratanam, 2000; Petrusa, 2002; Clauser and Schuwirth, 2002). These methods focus on maximum objectivity and standardised test conditions as prerequisites for reliable assessment. Despite the popularity of these types of assessment, 'in vivo' performance continues to be the primary basis for appraising clinical competence (Van der Vleuten et al., 2000; Williams et al., 2003). As a matter of fact, there are reasons to assume that practice-based assessment will occupy an increasingly prominent position in professional education. The increasing emphasis on outcome-based and competency-based education is likely to favour assessment methods that integrate relevant competencies (Van der Vleuten and Schuwirth, 2005). Assessment in authentic situations can focus on how students combine knowledge and skills, judgments and attitudes in dealing with realistic problems of professional practice. Moreover, on-going assessment of performance in day-to-day practice enables assessment of a range of essential competencies, some of which cannot be validly assessed otherwise, such as professional behaviour, efficient organisation of work, communication in teamwork, and continuous learning skills (McGaghie, 1993; Turnbull, 2002; Prescott et al., 2002). Therefore, in-training assessment (ITA), defined as multiple observations and assessment of performance in the setting of day-to-day practice, will remain an invaluable tool in comprehensive and valid assessment of clinical competence. However, while ITA may come closest to measuring habitual performance, the use of clinical performance ratings is not undisputed. Research has consistently shown considerable weaknesses, particularly regarding accuracy and reliability (Van der Vleuten et al., 2000; Sloan et al., 1995). For instance, in clerkship settings, assessors tend to give above average ratings which barely distinguish between students despite obvious differences in performance (e.g. Kwolek et al., 1997; Speer et al., 2000; Nahum, 2004). Furthermore, raters appear to use a 1 or 2 dimensional concept of performance and fail to distinguish between more detailed performance dimensions (Silber et al., 2004; Ramsey et al., 1993; Verhulst et al., 1986). Leniency, halo-effects and range restriction are much-discussed rater errors that are assumed to contribute to the inaccuracy of performance ratings. In addition, research has revealed a lack of rating consistency between raters and within raters, across different occasions, with reliability coefficients approaching zero (Van Barneveld, 2005; Gray, 1996; Noel et al., 1992; Littlefield et al., 1991). Finally, the validity of interpretations based on ITA scores is questionable due to content specificity, lack of discrimination between items and low correlations with other assessment formats (e.g. Kahn et al., 2001; Hull et al., 1995; Turnbull and van Barneveld, 2002).

Most criticisms of ITA stem from assessment views that are consistent with the quantitative psychometric framework. Central to this framework is judgment of performance through the inference of a *true* score, reflecting 'true' performance. Characteristics of the psychometric perspective are the pursuit of a specified level of consistency that is assumed to be conditional on technically sound measurement and the assumption of error when repeated measurements fail to yield consistent results. Raters are considered to be interchangeable ’measurement instruments’, and ratees’ ability is assumed to be a fixed, permanent and acontextual attribute. Rater effects, changes in context and interactions between ratees and tasks or contexts are regarded as unwanted sources of score variation (bias), compromising the utility of assessment results. Consequently, attempts to improve ITA from this perspective have focused primarily on standardisation and objectivity of measurement by adjusting assessment instruments, rating scale formats and enhancing raters’ accuracy and consistency through rater training. Despite these efforts, the general feeling prevails that clinical ratings have serious limitations (Williams et al., 2003) and that considerable improvement in assessing clinical competence is attainable by perfecting ITA (Turnbull and Cleveland, 2002; Holmboe, 2004).

The psychometric approach, however, may limit more meaningful educational approaches to ITA by its disregard for key issues within the mechanisms of the assessment process. The psychometric approach tends to ignore the role of the assessment context and issues of concern to those involved in the assessment task. ITA is predicated on active involvement of both students (ratees) and assessors (raters), who integrate personal goals, situational cues and organisational demands in the assessment process. For instance, in clinical education, ITA typically takes the form of assessment embedded in teaching, focusing on the assessment of individual students’ competence development. Learning processes and assessment tasks are largely determined by the dynamic context of patient care. The learning agenda is shaped by interactions among patients, teachers (assessors) and students, requiring teachers and students to negotiate assessment tasks and performance criteria. Also, if the essence of ITA is assessment of specific competencies that can only be validly assessed in real life practice, assessment tasks must intentionally capture *performance in context*. Essential aspects of authentic performance include students’ responses to particular factors that are determined by the context and time of the assessment. In other words, real life performance cannot be defined independently of the event and its context. As a consequence, consistency of measurement may be attained at a superficial level only, while the most critical aspects of performance will vary across contexts, time and individuals (Delandshere and Petrosky, 1998). Thus, performance ratings in ITA will inevitably reflect interpretations of students’ observed performance on poorly standardised tasks, related to individual learning goals, contextual factors and implied agreement on performance criteria.

Finally, ITA typically takes place in an organisational context which is determined by time constraints, competing goals (patient care, teaching, management) and a vague and frequently implicit set of norms and values in relation to performance assessment (Williams et al., 2003; Hoffman and Donaldson, 2004). Research into performance appraisal in various professional fields, including business and industry, the military and nursing, has indicated that it is not just student behaviour that is affected by contextual factors but rater behaviour as well (Murphy and Cleveland, 1995; Judge and Ferris, 1993).

On the basis of its educational and organisational characteristics, we argue that a new conceptual model of performance assessment is needed to make progress in research into and the use of ITA.

In the context-bound world of ITA, an exclusive focus on psychometric, context free criteria, such as accuracy or consistency of scores across tasks and raters, seems no longer appropriate. ITA involves judgement and decision making processes in which raters are no passive measurement instruments. On the contrary, raters are to be seen as active information processors in a complex social environment, continuously challenged to sample and make sense of performance data, and to judiciously use their personal judgements in public decision making (i.e. performance rating).

The role of raters as the key players in performance assessment lies at the centre of this paper. We will discuss three issues that we believe are crucial to performance assessment: raters' judgment and decision making processes from a social cognitive perspective; environmental factors that influence raters' motivation and goals; and the relationship between performance theories, values and beliefs and the practice of performance assessment.

Finally, we will present an alternative approach to performance assessment, based on insights from other disciplines and integrating perspectives from different theoretical frameworks. We favour a social psychological approach to performance assessment over the objective measurement perspective. In addition, we will address a number of implications of this approach for ITA practice and we will advocate directions for future research that are likely to advance our understanding of practice-based performance assessment.

Rater judgment and rater cognition

At the heart of ITA lie the cognitive processes and structures used by raters in forming impressions of and judging ratees' behaviour in the complex social environment of clinical practice. Research into social cognition has demonstrated that cognitive processes in judgment and decision making can be divided into a) relatively automatic top-down or schema-based information processing and b) relatively deliberate bottom-up or data-based information processing (Hogg, 2003; Hodgkinson, 2003; Fiske and Taylor, 1991).

In bottom-up processing, all the available information is attended to. Factual details and implications of observed behaviour are recalled, combined and weighted without reference to earlier experiences or prior knowledge. Research has shown, however, that individuals' limited ability for processing external stimuli favours the use of adaptive mechanisms, such as simplified representations of reality (knowledge structures referred to as categories, schemata or scripts) and mental shortcuts or heuristics in cognitive decision making (Komatsu, 1992). Once invoked by situational cues, these schematic knowledge structures and heuristics facilitate fast top-down processing, i.e. reliance on prior knowledge and preconceptions. In top-down processing, much of the information that might contribute to impression formation is lost and judgments are driven by global, holistic impressions. Generally, schematisation and automation are determined by formal and informal learning experiences and reflect the efficient and effective information processing of experts. Expert performance is characterised by attention to the most meaningful and relevant contextual events, effective encoding and retrieval of information, appropriate solutions and accuracy of judgment (Chi et al., 1989; Schmidt et al., 1990). Once established, schemata may be virtually impervious to change. Despite its obvious benefits, top-down information processing has the downside of potentially hampering the full use of all the information contained in situations in forming impressions and gaining an adequate understanding of the environment. Top-down information processing thus may encourage thinking in stereotypes, inaccurate filling of data gaps (typical but inaccurate information), rejection of relevant and possibly significant information and inhibition of disconfirmation of existing knowledge structures (Fiske and Taylor 1991; Walsh, 1995).

With this in mind, it may be informative to refer to similar concepts in research on clinical reasoning (Eva, 2004; Norman, 2005). This research attempts to understand reasoning strategies used by clinical experts – as compared to beginners in the field – when solving medical problems or making clinical decisions. Current research has shown that both analytical (bottom-up) and non-analytical (top-down) strategies are being used in clinical decision-making, at all levels of expertise. Research findings suggest that although the ability to successfully use non-analytic reasoning seems to increase with medical expertise, the approach is very effective in terms of diagnostic accuracy, even among novices. Research findings also show that reliance on non-analytical reasoning can be a source of diagnostic error. Therefore, optimal decision making most probably requires interactivity between non-analytical and analytical reasoning strategies (Eva, 2005).

Research into performance appraisal has suggested that both top-down and data-driven processing may occur, depending on contextual cues and the demands of the rating task. The objective of rating (i.e. establishing differences between or within persons), the rating-scale format and the complexity of the rating task affect which approach is used for information gathering, encoding and categorising, and thus the accessibility and availability of the information to be used in appraisal decisions (DeNisi and Williams, 1988; Jelley and Goffin, 2001).

Several studies have supported the existence of top-down or schema-driven processing in raters. For instance, research of performance rating in assessment centres and industrial organisations has shown that raters' recall of overall assessments is more accurate than that of supportive detail. Top-down ratings have been shown to be more accurate than data-driven ratings, with rating accuracy increasing with raters' experience and expertise (Lievens, 2001; Kozlowski and Mongillo 1992; Cardy et al., 1987). The findings that schemata are used to fill information gaps and direct raters' attention to schema-consistent information suggest that top-down information processing occurs, and may be responsible for one of the much-criticised traits of performance rating, namely the blurring of performance dimensions, or halo effects (Lievens, 2001; Lance et al., 1994, Fiske and Taylor, 1991; Zedeck, 1986).

Accuracy of rating is generally assumed to be a function of accuracy of observation and recall of behaviour. Consequently, bottom-up or data-driven processing in performance assessment is frequently stimulated by asking raters to observe and memorise those behaviours that have to be relied upon in generating performance ratings. However, as said earlier, research evidence on performance appraisal casts doubt on this relationship. Several findings have implied that raters with high accuracy at the behavioural level may make poor judgments. Similarly, raters who provide accurate holistic ratings, relying on on-line impressions, may show little accuracy in assessing behaviour - especially when the rating task is delayed (Murphy and Balzer, 1986, 1989). In other words, preserving all of the behavioural details is not a prerequisite for arriving at accurate judgments (Middendorf and Macan, 2002; Sanchez and DeLaTorre, 1996). Nevertheless, there is evidence that assessment features, such as behavioural checklists or structured diaries, may enhance the organisation of information in memory. These features may contribute to the usefulness of appraisal information for feedback purposes or raters' ability to substantiate their ratings, which is important to ensure fairness of assessment (DeNisi and Peters, 1996; Sanchez and DeLaTorre, 1996). A laboratory study by DeNisi et al. (1989) showed that diaries enhanced the accuracy of ratings and meaningful feedback by raters. This shows that, at least in some conditions, ratings may benefit from interventions that help raters organise and recall complex information. There is also evidence that for record keeping to be really effective it should follow closely on the observation of performance (Sanchez and DeLaTorre, 1996).

As the use of heuristics and schemata occurs relatively spontaneously, schema-based processing is likely to be a dominant automatic response in performance assessment in complex settings. The notion of schemata as simplified, persistent mental representations that enhance cognitive efficiency may offer a plausible explanation for some consistent findings in research on performance assessment, i.e. the limited success of rater training programmes and persistence of rater bias. Performance schemata are by definition idiosyncratic, representing unique individual experiences and understanding of performance. Obviously, raters' knowledge representations with regard to performance are bound to differ, depending on their professional experience and informal and formal communication (socialisation and training). Seeing that these unique cognitive schemata serve as organising frameworks for raters' cognitive processes in judgment and decision making, 'objective' situational stimuli may engender very different representations of reality, i.e. performance ratings will always reflect 'subjective' interpretations of situational behaviours. Thus the principles of schema-based reasoning in performance appraisal require taking account of raters' knowledge representations and performance theories as the primary basis for judgment and rating. We will elaborate on this later. Paying attention to cognitive approaches alone, however, does not suffice to achieve a full understanding of appraisal processes in the context of real life assessment. Recent research has shown that judgment and decision making are highly susceptible to mood and emotions (e.g., Forgas and George, 2001). Furthermore, there is increasing recognition of the importance of contextual factors. Different sources of rater motivation, including factors in the assessment context and affective states, may favour the use of different processing strategies, which affects the quality of ratings (Harris, 1994; Forgas, 2002). For example, the use of schemata will be stimulated when there are situational constraints (time pressures, distraction by other tasks) or with relatively low impact judgments (Siemer and Reisenzein, 1998; Rothman and Schwarz, 1998), whereas outcome dependency and accountability will tend to promote data-driven or bottom-up information processing (Fiske and Taylor, 1991). These findings have encouraged increased integration of cognitive, social, motivational and organisational perspectives in approaches to research of performance appraisal (Hodgkinson, 2003). In the next section, we will focus on contextual factors that have a potential impact on assessment outcomes by influencing raters' motivation and goals in ITA.

The educational and social context: raters' motivation and raters' goals

The impact of factors in the assessment context on students' learning behaviour has been examined in depth (Van Luijk et al., 1990; McIlroy et al., 2002; McDowell, 1995; Crooks, 1998). Findings from research into performance appraisal indicate that contextual factors also affect raters' behaviour and thus the quality of ratings (Judge and Ferris, 1993; Murphy and Cleveland, 1995; Williams et al., 2003; Hawe, 2003). This suggests that context may be a good starting point for examining components of the assessment process. Some critical contextual factors that mediate the relationship between raters and assessment are

- Purposes and use of assessment results;
- Consequences of ratings – rewards and threats; trust and accountability; and
- Organisational complexity, values and norms.

Purposes and use of assessment results

Assessment purposes may affect the quality and usefulness of rating outcomes in several ways. First of all, purpose may dictate the administrative features of rating, ranging from frequency to type of rating scale. For instance, for administrative purposes, it may be desirable to rely on infrequent and global ratings, whereas improvement of performance generally benefits from frequent feedback on specific performance dimensions.

As indicated before, an indirect effect of rating purposes may be mediated by basic cognitive processes of information acquisition, storage and recall (Landy and Farr, 1980; Murphy et al., 1984; Williams et al., 1985; Greguras, 2003). The potential of continuous performance assessment to serve both formative and summative purposes has prompted attempts to integrate these two functions based on considerations of efficiency and new approaches to assessment (e.g. Prescott, 2002). However, findings about the influence of assessment purposes on raters' cognitions raise doubts as to the effectiveness of combining summative and formative functions. For instance, faced with changes in rating purpose, raters have been found to have difficulty tailoring ratings to the new purpose; raters' information processing has been found to be impaired when the purpose of the final ratings differed from the purpose raters had in mind when observing ratees' performance (DeNisi and Williams, 1988). The observed impact of assessment purpose, makes it the more deplorable that research has suggested that organisations tend to convey vague or conflicting messages about rating purposes, especially when performance appraisal is used to serve multiple goals (Murphy and Cleveland, 1995; Cleveland et al.,1989). In ITA for instance, educational institutions may focus on the summative purpose of assessment (ranking of students), whereas supervisors (raters) may feel that its primary purpose should be to give feedback to students about the strengths and weaknesses of their performance. This may create a conflict of interest between rating purposes set by the organisation and raters' conceptions of their role in the organisation, i.e. that of mentors and coaches of students' competence development. Conflict may also arise in multi-purpose appraisal systems due to discrepancies between the actual use of assessment results and the assessment purposes communicated to raters. For instance, Hawe (2003) found that management who support the role of raters as the profession's gatekeepers (selection) may at the same time tell staff that there is concern about the retention of students and the related funding of the institution. This may lead to management challenging or even overturning low ratings, frustrating raters who were willing to provide accurate ratings. In these cases, raters have to weigh appraisal purposes, which may result in ratings that reflect ‘politically correct’ judgments rather than accurate performance appraisal (Mero and Motowidlo, 1995; Murphy and Cleveland, 1995). Finally, rating purpose may affect rater motivation, depending on the consequences of ratings as described below.

Consequences of ratings: rewards and threats; accountability and trust

A major complaint about performance ratings is that they tend to be inflated. Research findings indicate that leniency in performance rating is strongly affected by contextual effects on rater motivation. Perceived rewards, threats and accountability are major motivational pressures that influence raters’ behaviour. A framework for understanding leniency in performance appraisal was provided by Tetlock’s research (1983, 1985), as cited by Hauenstein (1992), in which the rating process is seen as ‘political’ decision making by raters who are accountable to others and motivated to seek the approval of those to whom they feel accountable. Raters will tend to bias their decisions towards what is acceptable to others. They may feel accountable to their supervisors (and management) and to ratees. When supervisors or management send strong messages that ratings must be justified, raters will be pressed to be thorough and careful in observing, recording and documenting performance and provide accurate ratings (Mero et al., 2003; Hauenstein, 1992). Similarly, a well designed assessment system (e.g. acceptance of rating forms, well defined performance dimensions and requirements of procedural justice) is likely to increase raters’ perceptions of accountability. By contrast, when raters perceive an increased accountability to ratees, they may be less motivated to provide accurate ratings and more likely to please ratees by giving lenient judgments (Klimoski and Inks, 1990). Unfortunately, most educational settings offer no (extrinsic) rewards for rating accuracy and accuracy may do no more than frustrate raters, as shown for example in the case study by Hawe (2003).

Accountability and assessment consequences are associated with raters’ trust in the assessment process. Trust reflects raters’ belief that consequences are fair and just and decision making is based on accurate ratings. Trust appears to influence the psychometric quality of ratings and acceptance of the rating system. A study by Bernardin and associates (1981) showed that trust in the appraisal process may account for 32% of the variance in ratings, with raters who feel a high degree of trust providing less lenient ratings. Good working relationships, well-defined roles, opportunity to observe ratees’ behaviour, high quality feedback (specific, honest) and low tolerance for political manipulation all contribute to trust in and trustworthiness of assessment (Murphy and Cleveland, 1995; Longenecker and Gioia, 2000; Piggot-Irvine, 2003).

Within the context of clinical education, several factors may contribute to leniency of performance ratings. Clinical supervisors often fulfil the dual roles of mentor-coach and assessor and may not be equipped to deal with these seemingly conflicting tasks. They may have difficulty giving students feedback about weaknesses in performance while maintaining a supportive student-supervisor relationship. Supervisors may be tempted to resort to upward distortion of ratings to avoid difficult feedback sessions and defensive reactions from students. Furthermore, extrapolating research findings about performance appraisal in the field of business and organisation, supervisors may not be very concerned with accuracy in performance rating

(Harris, 1994). Clinical supervisors' main concern may well be to establish and maintain high levels of student motivation. Although guidelines may emphasise accuracy of ratings, supervisors may believe that accurate low ratings will tend to turn into self-fulfilling prophecies by demotivating students who will subsequently perform at substandard levels. From this point of view, distortion of ratings may even be justified as being good teaching practice. Finally, both assessors and students tend to interpret performance assessment as at least in part a judgment of personal worth (Hawe, 2003). Supervisors may interpret assessments as reflecting on their competence as a teacher. Supervisors may feel accountable for poor student performance; they may feel they have failed as a teacher or fear that their competence as a teacher will be questioned. In these situations, supervisors will tend to bypass guidelines and assessment criteria in making decisions (Hawe, 2003; Harris, 1994).

Organisational complexity, norms and values

Research into performance appraisal in industrial organisations has identified several other categories of contextual factors that influence the rating process (Murphy and Cleveland, 1995). Rater behaviour will be affected by organisational norms and values regarding competence assessment, with low or below average ratings being unacceptable in some organisations, even if they are accurate. Research findings have also indicated that high entrance or educational standards for admission into a professional community may cause raters to be reluctant to assign average or low ratings. Although there is as yet no research evidence to back this up, implicit organisational norms and pressures for conformity may be a significant factor in inflating ratings in clinical settings. In addition, the fact that patient care is a team effort may complicate assessment of individual contributions. The increasing complexity of assessment tasks will affect raters' judgments and decision strategies, while time constraints and competing responsibilities may hamper careful information processing.

In summary, the evidence on appraisal processes from several domains suggests that contextual factors are key mechanisms, with important effects on raters' goals and motivation. Consequently, rating behaviour that is commonly labelled as rater error or inaccuracy may be attributable to (conscious or subconscious) raters adapting to situational cues, rewards or feedback.

Theories, values and beliefs about performance

Without clearly articulated and shared theories of performance, raters are unlikely to hold common definitions of work performance. Evidence from the field of performance appraisal has shown that framing performance assessment as a psychometric problem may lead to preoccupation with raters' errors and ways to eliminate them, while the development of organisation specific theories of work performance is neglected (Sulsky and Keown, 1999). An overview of research on assessment in medical education shows a similar preoccupation with assessment design, rater training and the development of rating scale formats in order to optimise the psychometric properties of performance assessment (Van der Vleuten and

Schuwirth, 2005). Assessment instruments often reflect designers' implicit theories about performance and describe performance dimensions and standards against which observed behaviour is to be measured. Although these external guidelines may be useful, they are generally unable to account for raters' judgments and decisions.

In most ITA approaches, rating scales reflect performance dimensions derived from formal job/task analyses, standards are defined in terms of more or less concrete behaviours or outcomes, and raters are expected to judge performance in terms of deviations from the concrete standard. However, as indicated before, judgments of real life performance in a social context will inevitably involve 'subjective' interpretation of 'objective' information, matching internal concepts of performance, which are rooted in experience and training. Raters' abstract task schemata as well as their perceptions of the purposes and consequences of the appraisal process will determine which standards they actually use. These 'intuitive' standards may vary considerably from judge to judge and disagreement between raters may say more about differences in task schemata or interpretation of assessors' tasks than about differences in perceptions of ratees' behaviour. Research on performance appraisal also suggests that the dimensions that supervisors emphasise do not necessarily match the dimensions derived through formal job analysis. In all probability, job performance refers to two distinct sets of behaviours: those contained in formal job descriptions and those defined by the social context of work and organisations – i.e. contextual performance or extra-role behaviours (Borman and Motowidlo, 1997; Johnson, 2001). Extra-role behaviours contribute to the organisational, social and psychological working environment and are typically context bound. Research findings have shown that these behaviours strongly influence raters' search strategies and ratings by supervisors. For instance, supervisors generally place high importance on persisting with enthusiasm and extra effort, volunteering to carry out extra duties, job-task conscientiousness, contributions to a more positive working climate and handling work stress. These traits are often considered to be more important than specific aspects of task proficiency (Johnson, 2001). Raters also tend to include situational constraints in their performance assessment. Situational variables that enhance or depress performance are weighted and factored into assessments (Sulsky and Keown, 1999). Individual ratees' competence development may also influence raters' perceptions of the relative importance of performance dimensions. For instance, when performance components that tend to be predicted by cognitive ability have become automatic, their contribution to overall performance ratings may decrease (Govaerts et al., 2005). Finally, judgments about performance involve values as well as objective information. Different judges are likely to hold different values, particularly when they have different experiences, prior knowledge or occupational backgrounds. Disagreement on performance assessment may thus stem either from disagreement on facts or disagreement on values. In short, raters will judge observed behaviour against personal values and internal performance theories as well as external guidelines, using implicit internal standards to assess performance. Whether raters will comply with the standards defined by the assessment designers will depend on the extent to which raters can identify with and have internalised these external directions These processes are subject to a broad range of influences.

It is our view that lack of convergence across rating sources reflects the complexity and context-specific nature of performance assessment, the lack of explicit and unifying performance theories and raters' use of idiosyncratic theoretical frameworks. From a traditional psychometric perspective, this raises the problem of unwanted measurement bias. However, several researchers in the domain of performance appraisal have proposed an alternative view. They argue that different measurement sources may result in multiple true performance scores, each capturing both common and unique aspects of ratees' performance. Raters from different perspectives may rate differently because they observe different aspects of performance, and differences in ratings may very well reflect true differences in performance (Lance et al., 1992). This suggests that there is room for honest disagreement and performance ratings from different sources "may be equally valid, even though not highly correlated" (Landy and Farr, 1980, p.76). Indeed, the concept of a true score has been challenged within traditional assessment approaches as well. For instance, in the field of standard setting, the belief in a true cut score has been replaced by the knowledge that all standard setting depends on subjective judgments, i.e. the values and beliefs of the people constructing the cut scores (Zieky, 2001 p. 45).

The measurement approach to performance assessment is grounded in the assumption that students will perform consistently across tasks and performance dimensions. However, given the complexity of assessment of real-life performance, the validity of these performance theories seems questionable. Nichols and Smith (1998) have argued that a more appropriate theoretical framework for assessing complex performance would be one that allows for different ratees using different procedures and strategies across different tasks and occasions. Such a framework would rely on the assumption that ratees are active learners who construct their own knowledge on the basis of unique practical experiences. It would be consistent with the view that performance is highly contextual and definitions of 'good performance' may encompass many different approaches to assessment tasks. Moreover, this framework would link performance interpretation to theories of learning and problem solving (Nichols and Smith, 1998). When this framework is applied to performance rating, a picture emerges of performance assessment as context specific and based on raters' unique cognitive structures. Therefore, in our view, a constructivist approach to assessment is equally applicable to raters as it is to ratees.

Concluding remarks

We have addressed several issues that we believe are critical in practice-based assessment, like ITA. Although a comprehensive description of performance assessment should also address the impact of rater-ratee and ratee-context interactions on the rating process, we have deliberately focused on the role of raters and the impact of context on raters' judgment and decision making processes. The reason for this is that these aspects have remained somewhat underexposed in the literature on assessment in medical education.

A constructivist, social-psychological approach to performance appraisal

We believe that the issues raised in this article make out a case for an approach to ITA that: is based on insights from other disciplines (Murphy and Cleveland, 1995), takes a predominantly constructivist, social-psychological perspective, and integrates elements of theories of cognition, motivation and decision making. A central assumption in the proposed approach is that ITA is a judgment and decision making process in which raters' behaviour is shaped by interactions between individuals and social context in which assessment occurs. Raters are no longer seen as passive measurement instruments, but as active information processors, who interpret and construct their personal reality of the assessment context. They gather information about ratees' performance and make public decisions (ratings) based on their personal evaluation (judgment) of that information. Situational cues, conceptions of good or poor performance, ratees' behaviour as well as cognitive and motivational rater variables, all contribute to rating outcomes. We take the view that raters' behaviour is motivated and goal directed, defined by raters' perceptions of the assessment system and its intended or unintended (negative) effects. Actual public ratings communicate raters' goals to other parties involved in the assessment process and they may differ from raters' personal judgments or feedback to ratees (Murphy and Cleveland, 1995; Murphy et al., 2004). This implies that real-life performance assessment is less about measurement and more about reasoning, problem solving and decision making in a dynamic environment, akin to clinical reasoning and decision making in medical practice (Norman, 2005).

Implications of the new approach

Some of the implications of the proposed approach for current ITA practice will be discussed briefly.

From the perspective of our approach, it does not make sense to exclusively attribute raters' errors to raters' inability to produce accurate ratings. Raters are no passive measurement instruments, and they should not be treated as such. Discrepancies between actual performance and ratings may simply reflect effects of forces that discourage accurate rating, and failure to discriminate between persons or dimensions may constitute adaptive behaviour. Depending on their goals in the rating process, raters may want to enhance the usefulness of performance appraisal through 'motivated distortion' of ratings. In fact, raters' behaviour may be driven more strongly by situational variables than by actual differences between ratee variables (Murphy and Cleveland, 1995). It may, therefore, be a more effective strategy for improving ITA practice to focus on the assessment context than to focus on individual raters.

In this respect, several factors in the assessment context deserve special attention:

1. Trust in and acceptance of the assessment system by raters and ratees is a crucial factor. The concept of trust is related to the concept of consequential validity. Prerequisites for trust in the assessment system are well documented and include authenticity, fairness, honesty, transparency of procedures (due process), well-defined roles and high quality feedback (Messick, 1994; Piggot-Irvine, 2003; Taylor et al., 1995; Erdogan et al., 2001). Acceptance of performance appraisal may be enhanced by involving raters in assessment development and by participation of ratees in the assessment process. In addition, the organisation should

create conditions that allow raters to be thorough and honest in their assessment. This implies that raters should have adequate opportunities to gather and document relevant information and feel confident about providing performance ratings and feedback (rater training). Raters should be motivated to be careful and thorough, i.e. they should be accountable for the ratings they provide to both ratees and management. This requires documentation of evidence (direct observation, immediately followed by documentation of performance interpretations and feedback) as well as procedures that ensure interaction between the parties involved in the assessment process. Finally, management must provide clarity about assessment purposes and the use of assessment results. This requires not only transparent procedures but also open and honest communication about what is expected from all parties involved.

2. The underlying performance theories should be explicated and communicated to all parties involved in the assessment. However, it should be acknowledged that these performance theories and standards are in accordance with the values and beliefs of those who designed the assessment system. ’Truth’ will always be a matter of consensus (Johnston, 2004). Furthermore, interpretations of observed behaviour will always reflect raters’ personal performance theories, based on individual experiences, values, demands of the local context, and role perceptions, - as well as external guidelines. Inherent to the constructivist approach of our approach and the inevitability of personal performance constructs in assessment is the acceptance of different performance interpretations, i.e. multiple true performance scores. We think that this constructivist approach to assessment has important implications for assessment procedures (Krefting, 1991; Tigelaar et al., 2005; Rust et al., 2005). For instance, final decisions based on performance ratings should always incorporate input from many different rating sources. This is consistent with traditional approaches. However, within our approach, the main purpose of combining input from different rating sources is not to reach one ideal, “objective” decision through consensus on a mean effectiveness rating. The rationale for combining sources in the new model is that multiple interpretations of ratees’ performance may be equally valid and together present a rich and detailed report of competencies and situation-specific behaviours. This is in line with new developments in competence-based education and other forms of practice-based assessment, such as portfolios. It should be noted that we fully agree with Johnston (2004) that this does not imply that “any interpretation is acceptable, that ‘anything goes’”. Essentially, final decision making requires professional judgments that should be corroborated, motivated and substantiated in such a way that the decision is defensible and credible. Our argument is that discussing individual assessment processes, values and standards will contribute to a shared view about what is really important and constitutes meaningful performance assessment. In this process, organisational guidelines may be the starting point to create locally accepted appraisal systems that express the values and judgments of both the organisation and the assessors. Examples of such assessment practices can be found in the literature about portfolio assessment (Johnston, 2004; Delandshere and Petroski, 1998) and similar approaches have been developed in medical education (Pangaro, 2000; Schwind et al., 2004).
3. Traditionally, rater training has focused on acquainting raters with the assessment system and instruments and the skills to use them effectively. Most successful training methods

involve frame-of-reference training or cognitive modelling principles (Woehr and Hufcutt, 1994; Schleicher and Day, 1998; Lievens, 2001). However, the alternative approach to performance assessment which we propound in this paper offers several other perspectives on rater training. From the perspective of performance rating as goal-directed behaviour, rater training should focus not only on rater ability, but also (and perhaps even more so) on rater motivation. This implies that rater training should include awareness raising of internal values and beliefs about performance appraisal, accountability, potential role conflicts, feedback and skills for establishing trusting, open and non-defensive yet problem-confronting relationships.

4. Finally, the proposed assessment approach requires reappraisal of the framework for evaluating assessment methods. The psychometric framework may no longer be appropriate to exclusively evaluate assessment quality and we may need alternative criteria that are in line with the constructivist assessment approach. We need criteria that ensure rigour, without sacrificing the unique benefits of a more descriptive, qualitative approach. The Guba and Lincoln model for constructivist assessment may prove very useful in this context (Guba and Lincoln, 1989). This model has been recommended for other context-bound assessment programmes, for example in teacher education (Driessen et al., 2005; Tigelaar et al., 2005). Guba and Lincoln have described a number of criteria for assessing the quality of assessment, focusing on trustworthiness and authenticity of the assessment process. These criteria are partly incorporated in assumptions that underlie traditional approaches (parallel to validity, reliability and objectivity), but they also include fairness, shared understanding, individual constructions and learning, and intended consequences.

Implications for further research

This overview of research on performance assessment is by no means comprehensive and future research will have to confirm the adequacy of our model within the context of medical education. Rater cognition, rater motivation and rater training are important areas for research. This research should focus on the cognitive processes in appraisal and how these are affected by appraisal purposes, rater motivation and features of the assessment system. For instance, although new developments in assessment increasingly focus on integrating assessment and instruction, research findings seem to indicate that summative and formative purposes are incompatible in performance appraisal. Cognitive research seems to indicate that accuracy of rating relies on top-down processing, resulting in holistic performance assessments that disregard detailed information about performance dimensions. However, accuracy of rating reflects only one perspective on the usefulness of appraisal information. Records of behaviours may be needed to enable effective feedback, i.e. to identify strengths and weaknesses, as well as to substantiate judgments. This means that it is important to determine whether raters' cognitive limitations preclude combining formative and summative purposes in a single assessment system. If so, we should find ways to compensate for these limitations by modifying assessment instruments and procedures.

Features of the assessment system, such as performance dimensions and rating scale formats are important stimuli in performance appraisal. Research on rating scale formats has largely ignored cognitive issues. Different scales may tap into different cognitive processes, which may affect rating outcomes. A better understanding of these cognitive processes, the presence and the actual use of schemata in performance assessment may reveal a need for different approaches to designing assessment instruments and procedures.

There is also a need for research on performance theories in medical practice. A better understanding of raters' implicit performance theories, in particular, would increase our insight into performance judgments. When do raters use their own standards, when do they comply with external guidelines? Research should also focus on the categorisation processes that underlie performance schemata and are used to interpret and judge observed behaviour. More insight is needed into how raters combine and weigh different kinds of information. How are situational constraints factored into assessments, how do raters deal with inconsistencies between process and result?

Finally, research should address rater motivation and factors that influence rater motivation and rater goals. For instance, it is important to gain more insight into raters' perceptions of assessment systems and assessment purposes. What are implicit rater goals and how do they affect rating outcomes? Which context variables are important in encouraging (or discouraging) raters to provide high quality performance assessment, and how can these factors be influenced? What are efficient and effective ways to involve raters and ratees in the appraisal process? How can we achieve accountability and trust in the assessment system with limited resources and while maintaining feasibility and flexibility?

References

Barneveld van, C. (2005). The dependability of medical students' performance ratings as documented on in-training evaluations. *Academic Medicine, 80*(3), 309-312.

Bernardin, H.J., Orban, J.A. & Carlyle, J.J. (1981). Performance ratings as a function of trust in appraisal and rater individual differences. *Academy of Management Proceedings*, pp. 311-315.

Borman, W.C., Motowidlo, S.J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10*, 99-109.

Cardy, R.L., Bernardin, H.J., Abbott, J.G., Senderak, M.P. & Taylor, K. (1987). The effects of individual performance schemata and dimension familiarization on rating accuracy. *Journal of Occupational Psychology, 60*, 197-205.

Chi, M.T.H., Glaser, R. & Farr, M.J. (1989). *The Nature of Expertise*. Hillsdale: New Jersey.

Clauser, B.E. & Schuwirth, L.W.T. (2002). The use of computers in assessment. In: G.R. Norman, C.P.M. van der Vleuten & D.I. Newble (Eds.), *International Handbook of Research in Medical Education* (pp.757-792). Dordrecht: Kluwer Academic Publishers.

Cleveland, J.N., Murphy, K.R. & Williams, R.E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74*, 130-135.

Crooks, T. (1998). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438-481.

Delandshere, G. & Petrosky, A.R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher, 27*(2), 14-24.

DeNisi A.S. & Peters, L.H. (1996). Organization of information in memory and the performance appraisal process: Evidence from the field. *Journal of Applied Psychology, 81*(6), 717-737.

DeNisi, A.S., Robbins, & T. Cafferty, T.P. (1989). Organization of information used for performance appraisals: Role of diary-Keeping. *Journal of Applied Psychology, 74*(1), 124-129.

DeNisi, A.S & Williams, K.J. (1988). Cognitive approaches to performance appraisal. In: G. Ferris & K. Rowland (Eds.), *Research in Personnel and Human Resource Management* (Vol. 6). Greenwich, CT: JAI Press.

Driessen, E., Vleuten van der, C., Schuwirth, L., Tartwijk van, J. & Vermunt, J. (2005). The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: A case study. *Medical Education*, *39,* 214-220.

Erdogan, B., Kraimer, M.L. & Liden, R.C. (2001). Procedural justice as a two-dimensional construct. An examination in the performance appraisal context. *Journal of Applied Behavioural Science, 37*(2), 205-222.

Fiske, S.T. & Taylor, S.E. (1991). *Social cognition* (2nd ed). New York: McGraw-Hill.

Forgas, J.P. & George, J.M. (2001). Affective influences on judgments and behavior in organizations: An information processing perspective. *Organizational Behavior and Human Decision Processes, 86*(1), 3-34.

Forgas, J.P. (2002). Feeling and doing: Influences on interpersonal behavior. *Psychological Inquiry, 13*(1), 1-28.

Govaerts, M.J.B., Vleuten van der, C.P.M., Schuwirth, L.W.T. & Muijtjens, A.M.M. (2005). The use of observational diaries in in-training evaluation: Student perceptions. *Advances in Health Sciences Education,10*(3), 171-188.

Gray, J.D. (1996). Global rating scales in residency education. *Academic Medicine, 71*(1), S55-S63.

Greguras, G.J., Robie, C., Schleicher, D.J. & Goff III, M. (2003). A field study of the effects of rating purpose on the quality of multisource ratings. *Personnel Psychology, 56*, 1-20.

Guba, E. & Lincoln, Y. (1989). *Fourth generation evaluation*. London: Sage Publications.

Harris, M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management, 20*(4), 737-756.

Hauenstein, N.M.A. (1992). An information-processing approach to leniency in performance judgments. *Journal of Applied Psychology, 77*(4), 485-493.

Hawe, E. (2003). It's pretty difficult to fail: The reluctance of lecturers to award a failing grade. *Assessment and Evaluation in Higher Education, 28*(4), 371-382.

Hodgkinson, G.P. (2003). The interface of cognitive and industrial, work and organizational psychology. *Journal of Occupational and Organizational Psychology, 76*, 1-25.

Hoffman, K.G. & Donaldson, J.F. (2004). Contextual tensions of the clinical environment and their influence on teaching and learning. *Medical Education, 38*, 448-454.

Hogg, M.A. (2003). Introducing social psychology. In: Hogg, M.A. (Ed.), *Social Psychology, Vol. I: Social Cognition and Social Perception.* pp xxi-lix. London: Sage Publications

Holmboe, E.S. (2004). Faculty and the observation of trainees' clinical skills: Problems and opportunities. *Academic Medicine, 79*(1), 16-22.

Hull, A.L., Hodder, S., Berger, B., Ginsberg, D., Lindheim, N., Quan, J. & Kleinhenz, M. (1995). Validity of three clinical performance assessments of internal medicine clerks. *Academic Medicine, 70*(6), 517-522.

Jelley, R.B. & Goffin, R.D. (2001). Can performance-feedback accuracy be improved? Effects of rater priming and rating-scale format on rating accuracy. *Journal of Applied Psychology, 86*(1), 134-144.

Johnson, J.W. (2001). The relative importance of task and contextual performance dimensions to supervisor judgements of overall performance. *Journal of Applied Psychology, 86*(5), 984-996.

Johnston, B. (2004). Summative assessment of portfolios: An examination of different approaches to agreement over outcomes. *Studies in Higher Education, 29*(3), 395-412.

Judge, T.A. & Ferris, G.R. (1993). Social context of performance evaluation decisions. *Academy of Management Journal, 36*(1), 80-105.

Kahn, M.J., Merrill, W.W., Anderson, D.S. & Szerlip, H.M. (2001). Residency program director evaluations do not correlate with performance on a required 4th-year Objective Structured Clinical Examination. *Teaching and Learning in Medicine, 13*(1), 9-12.

Klimoski, R. & Inks, L. (1990). Accountability forces in performance appraisal. *Organizational Behavior and Human Decision Processes*, *45*, 194-208.

Krefting, L. (1991). Rigor in qualitative research: The assessment of trustworthiness. *American Journal of Occupational Therapy, 45*, 214-222.

Komatsu, L.K. (1992). Recent views on conceptual structure. *Psychological Bulletin, 112*(3), 500-526.

Kozlowski, S.W.J. & Mongillo, M. (1992). The nature of conceptual similarity schemata: Examination of some basic assumptions. *Personality and Social Psychology Bulletin, 18*, 88-95.

Kwolek, C.J., Donnelly, M.B., Sloan, D.A., Birrell, S.N., Strodel, W.E. & Schwartz, R.W. (1997). Ward evaluations: Should they be abandoned? *Journal of Surgical Research*, *69*(1), 1-6.

Lance C.E., LaPointe J.A. & Stewart, A.M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology, 79*(3), 332-340.

Lance, C.E., Teachout, M.S. & Donnelly, T.M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *J Appl Psych, 77*(4), 437-452.

Landy F.J. & Farr, J.L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability and discriminant validity. *Journal of Applied Psychology, 86*(2), 225-264.

Littlefield, J.H., DaRosa, D.A., Anderson, K.D., Bell, R.M., Nicholas, G.G. & Wolfson, P.J. (1991). Assessing performance in clerkships: Accuracy of surgery clerkship performance raters. *Academic Medicine, 66*(9), S16-S18.

Longenecker, C.O. & Gioia, D.A. (2000). Confronting the "politics" in performance appraisal. *Business Forum*, *25*(3,4), 17-23.

Luijk van, S.J., Van der Vleuten, C.P.M. & Schelven, R.M. (1990). The relation between content and psychometric characteristics in performance-based testing. In W. Bender, R.J. Hiemstra, Scherpbier. A.J.J.A. & Zwierstra R. P. (Eds.), *Teaching and Assessing Clinical Competence* (pp.497-502). Groningen: Boekwerk Publications.

McDowell, L. (1995). The impact of innovative assessment on student learning. *Innovations in Education and Training International, 32*(4), 302-313.

McGaghie, W.C. (1993). Evaluating competence for professional practice. In: L. Curry L, J.F.Wergin and Associates (Eds.), *Educating professionals: Responding to new expectations for competence and accountability* (pp. 229-261). San Francisco: Jossey-Bass Inc., Publishers.

McIlroy, J.H., Hodges, B., McNaughton, N. & Regehr, G. (2002). The effect of candidates' perceptions of the evaluation method on reliability of checklist and global rating scores in an Objective Structured Clinical Examination. *Academic Medicine, 77*, 725-728.

Mero, N.P. & Motowidlo, S.J. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology, 80*(4), 517-524.

Mero, N.P., Motowidlo, S.J. & Anna, A.L. (2003). Effects of accountability on rating behavior and rater accuracy. *Journal of Applied Social Psychology, 33*(12), 2493-2514.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Middendorf, C.H. & Macan, T.H. (2002). Note-taking in the employment interview: Effects on recall and judgments. *Journal of Applied Psychology, 87*(2), 293-303.

Murphy K.R. & Balzer, W.K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluation: Consequences for rating accuracy. *Journal of Applied Psychology, 71*, 39-44.

Murphy, K.R. & Balzer, W.K. (1989). Rating errors and rating accuracy. *Journal of Applied Psychology*, *74*(4), 619-624.

Murphy, K.R. & Cleveland, J.N. (1995). *Understanding performance appraisal. Social, Organizational and Goal-based Perspectives*. Thousand Oaks, CA: Sage Publications.

Murphy, K.R., Cleveland, J.N., Skattebo, A.L., & Kinney, T.B. (2004). Raters who pursue different goals give different ratings*. Journal of Applied Psychology, 89*(1), 158-164.

Murphy, K.R., Balzer, W.K., Kellam, K.L. & Armstrong, J. (1984). Effects of purpose of rating on accuracy in observing teacher behavior and evaluating teaching behavior. *Journal of Educational Psychology, 76*, 45-54.

Nahum, G.G. (2004). Evaluating medical student obstetrics and gynecology clerkship performance: Which assessment tools are most reliable? *American Journal of Obstetrics and Gynaecology, 191*, 1762-71.

Nichols, P.D. & Smith, P.L. (1998). Contextualizing the interpretation of reliability data. *Educational Measurement: Issues and Practice, 17*, 24-36.

Noel, G.L., Herbers, J.E.J., Caplow, M.P., Cooper, G.S., Pangaro, L.N. & Harvey, J. (1992). How well do internal medicine faculty members evaluate the clinical skills of residents? *Annals of Internal Medicine, 117*, 757-65.

Norman, G. (2005). Research in clinical reasoning: Past history and current trends. *Medical Education, 39*(4), 418-427.

Pangaro, L.N. (2000). Investing in descriptive evaluation: A vision for the future of assessment. *Medical Teacher, 22*(5), 478-481.

Petrusa, E.R. (2002). Clinical performance assessments. In: G.R. Norman, C.P.M. van der Vleuten & D.I. Newble (Eds.), *International Handbook of Research in Medical Education* (pp.673-709). Dordrecht: Kluwer Academic Publishers.

Piggot-Irvine, E. (2003). Key features of appraisal effectiveness. *The International Journal of Educational Management, 17*(4), 170-178.

Prescott, L.E., Norcini, J.J., McKinlay, P. & Rennie, J.S. (2002). Facing the challenges of competency-based assessment of postgraduate dental training: Longitudinal Evaluation of Performance (LEP). *Medical Education, 36*, 92-97.

Ramsey, P.G., Wenrich, M.D., Carline, J.D., Inui, T.S., Larson, E.B. & Logerfo, J.P. (1993). Use of peer ratings to evaluate physician performance. *Journal of the American Medical Association, 269*(13), 1655-1660.

Reznick, R.K. & Rajaratanam, K. (2000). Performance-based assessment. In: L.H. Distlehorst, G.L. Dunnington and J.R. Folse (Eds.), *Teaching and Learning in Medical and Surgical Education. Lessons learned for the 21st century* (pp. 237-243). Mahwah NJ: Lawrence Erlbaum Ass.

Rothman, A.J. & Schwarz, N. (1998). Constructing perceptions of vulnerability: Personal relevance and the use of experiential information in health judgments. *Personality and Social Psychology Bulletin, 24*(10), 1053-1064.

Rust, C., O'Donovan, B. & Price, M. (2005). A social constructivist assessment process model: How the research literature shows us this could be best practice. *Assessment & Evaluation in Higher Education, 30*(3), 231-240.

Sanchez, J.I. & DeLaTorre, P. (1996). A second look at the relationship between rating and behavioral accuracy in performance appraisal. *Journal of Applied Psychology, 81*(1), 3-10.

Schleicher D.J. & Day, D.V. (1998). A cognitive evaluation of frame-of-reference rater training: Content and process issues. *Organizational Behaviour and Human Decision Processes, 73*(1), 76-101.

Schmidt, H.G., Norman, G.R. & Boshuizen, H.P.A. (1990). A cognitive perspective on medical expertise: Theory and implications. *Academic Medicine, 65*(10), 611-621.

Schwind, C.J., Williams, R.G., Boehler, M.L. & Dunnington, G.L. (2004). Do individual attending post-rotation performance ratings detect resident clinical performance deficiencies? *Academic Medicine, 79*, 453-457.

Siemer, M. & Reisenzein, R. (1998). Effects of mood on evaluative judgements: Influence of reduced processing capacity and mood salience. *Cognition and Emotion, 12*(6), 783-805.

Silber, C.G., Nasca, T.J., Paskin, D.L., Eiger, G., Robeson, M. & Veloski, J.J. (2004). Do global rating forms enable program directors to assess the ACGME Competencies? *Academic Medicine, 79*, 549-556.

Sloan, D.A., Donnelly, M.B., Drake, D.B. & Schwartz, R.W. (1995). Faculty sensitivity in detecting medical students' clinical competence. *Medical Teacher, 17*(3), 335-342.

Speer, A.J., Soloman, D.J, & Fincher, R.M. (2000). Grade inflation in internal medicine clerkships: Results of a national survey. *Teaching and Learning in Medicine, 12*, 112-116.

Sulsky L.M. & Keown, J.L. (1999). Performance appraisal in the changing world of work: Implications for the meaning and measurement of work performance. *Canadian Psychology*, *39*, 1-2, 52-59.

Taylor, M.S., Tracy, K.B., Renard, M.K., Harrison, J.K. & Carroll, S.J. (1995). Due process in performance appraisal: A quasi-experiment in procedural justice. *Administrative Science Quarterly, 40*, 495-523.

Tetlock, P.E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology, 45*, 74-83.

Tetlock, P.E. (1985). Accountability: The neglected social context of judgment and choice. In: L.L. Cummings & B.M. Staw (Eds.), *Research in Organizational Behavior Vol 7* (pp 297-332). Greenwich, CT: JAI Press.

Tigelaar, D.E.H., Dolmans, D.H.J.M., Wolfhagen, I.H.A.P. & Vleuten van der, C.P.M. (2005). Quality issues in judging portfolios: Implications for organizing teaching portfolio assessment procedures. *Studies in Higher Education, 30*(5), 595-610.

Turnbull, J. & van Barneveld, C. (2002). Assessment of clinical performance: In-training evaluation. In: G.R. Norman, C.P.M. van der Vleuten & D.I. Newble (Eds.), *International Handbook of Research in Medical Education* (pp.793-810). Dordrecht: Kluwer Academic Publishers.

Verhulst, S., Colliver, J., Paiva, R. & Williams, R.G. (1986). A factor analysis of performance of first-year residents. *Journal of Medical Education, 61*, 132-34.

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

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

Vleuten van der, C.P.M., Scherpbier, A.J.J.A., Dolmans, D.H.J.M., Schuwirth, L.W.T., Verwijnen, G.M. & Wolfhagen, H.A.P. (2000). Clerkship assessment assessed. *Medical Teacher, 22*(6), 592-600.

Walsh J.P. (1995). Managerial and organizational cognition: Notes from a trip down memory lane. *Organization Science, 6*(3), 280-321.

Williams, K.J., DeNisi A.S., Blencoe, A.G. & Cafferty, T.P. (1985). The role of appraisal purpose: Effects of purpose on information acquisition and utilization. *Organizational Behavior and Human Performance, 35*, 314-339.

Williams, R.G., Klamen, D.A. & McGaghie, W.C. (2003). Cognitive, social and environmental sources of bias in clinical performance settings. *Teaching and Learning in Medicine, 15*(4), 270-292.

Woehr, D.J. & Huffcutt, A.I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organisational Psychology, 67*, 189-205.

Zedeck, S. (1986). A process analysis of the assessment centre method. *Research in Organizational Behavior, 8*, 259-296.

Zieky, M.J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G.J. Cizek (Ed.), *Setting Performance Standard: Concepts, methods and perspectives,* pp. 19-53. Mahwah NJ: Lawrence Erlbaum Associates.

CHAPTER 6

Workplace-based assessment: effects of rater expertise

published in Advances in Health Sciences Education 2011,
Vol 16, No 2: 151-165[3]

Marjan J.B. Govaerts
Lambert W.T. Schuwirth
Cees P.M. van der Vleuten
Arno M.M. Muijtjens

[3] Based on the paper entitled "Assessing Raters for Workplace-Based Assessment: A Cognition-Based Approach", presented at the AERA-conference, 2010, Denver USA. Awarded 'Best paper by New Investigator Division I'.

Abstract

Traditional psychometric approaches towards assessment tend to focus exclusively on quantitative properties of assessment outcomes. This may limit more meaningful educational approaches towards workplace-based assessment (WBA). Cognition-based models of WBA argue that assessment outcomes are determined by cognitive processes by raters which are very similar to reasoning, judgment and decision making in professional domains such as medicine. The present study explores cognitive processes that underlie judgment and decision making by raters when observing performance in the clinical workplace. It specifically focuses on how differences in rating experience influence information processing by raters.

Verbal protocol analysis was used to investigate how experienced and non-experienced raters select and use observational data to arrive at judgments and decisions about trainees' performance in the clinical workplace. Differences between experienced and non-experienced raters were assessed with respect to time spent on information analysis and representation of trainee performance; performance scores; and information processing –using qualitative-based quantitative analysis of verbal data.

Results showed expert-novice differences in time needed for representation of trainee performance, depending on complexity of the rating task. Experts paid more attention to situation-specific cues in the assessment context and they generated (significantly) more interpretations and fewer literal descriptions of observed behaviours. There were no significant differences in rating scores. Overall, our findings seemed to be consistent with other findings on expertise research, supporting theories underlying cognition-based models of assessment in the clinical workplace. Implications for WBA are discussed.

Introduction

Recent developments in the continuum of medical education reveal increasing interest in performance assessment, or workplace-based assessment (WBA) of professional competence. In outcome-based or competency-based training programs, assessment of performance in the workplace is a sine qua non (Van der Vleuten and Schuwirth, 2005). Furthermore, the call for excellence in professional services and the increased emphasis on life-long learning require professionals to evaluate, improve and provide evidence of day-to-day performance throughout their careers. Workplace-based assessment (WBA) is therefore likely to become an essential part of both licensure and (re)certification procedures, in health care just as in other professional domains such as aviation, the military and business (Cunnington and Southgate, 2002; Norcini, 2005).

Research into WBA typically takes the psychometric perspective, focusing on *quality of measurement*. Norcini (2005), for instance, points to threats to reliability and validity from uncontrollable variables, such as patient mix, case difficulty and patient numbers. Other studies show that the utility of assessment results is compromised by low inter-rater reliability and rater effects such as halo, leniency or range restriction (Kreiter and Ferguson, 2001; Van Barneveld, 2005; Gray, 1996; Silber et al., 2004; Williams and Dunnington, 2004; Williams et al., 2003). As a consequence, attempts to improve WBA typically focus on standardization and objectivity of measurement by adjusting rating scale formats and eliminating rater errors through rater training. Such measures have met with mixed success at best (Williams et al., 2003).

One might question, however, whether an exclusive focus on the traditional psychometric framework, which focuses on quantitative assessment outcomes, is appropriate in WBA-research. Research in industrial psychology demonstrates that assessment of performance in the workplace is a complex task which is defined by a set of interrelated processes. Workplace-based assessment relies on judgments by professionals, who typically have to perform their rating tasks in a context of time pressure, non-standardized assessment tasks and ill-defined or competing goals (Murphy and Cleveland, 1995). Findings from research into performance appraisal also indicate that contextual factors affect rater behaviour and thus rating outcomes (Levy and Williams, 2004; Hawe, 2003). Raters are thus continuously challenged to sample performance data; interpret findings; identify and define assessment criteria; and translate private judgments into sound (acceptable) decisions. Perhaps performance rating in the workplace is not so much about 'measurement' as it is about 'reasoning', 'judgment' and 'decision making' in a dynamic environment. From this perspective, our efforts to optimize WBA may benefit from a better understanding of raters' reasoning and decision making strategies. This implies that new and alternative approaches should be used to investigate assessment processes, with a shift in focus from quantitative properties of rating scores towards analysis of the cognitive processes that raters are engaged in when assessing performance.

The idea of raters as information processors is central to cognition-based models of performance assessment (Feldman, 1981; De Nisi, 1996). Basically, these models assume that rating outcomes vary, depending on how raters recognize and select relevant information (information acquisition); interpret and organize information in memory (cognitive representation of ratee behaviour); search for additional information; and finally retrieve and integrate relevant information in judgment and decision making. These basic cognitive processes are similar to information processing as described in various professional domains, such as management, aviation, the military and medicine (Walsh, 1995; Ross et al., 2006; Gruppen and Frohna, 2002). Research findings from various disciplines show that large individual variations in information processing can occur, related to affect, motivation, time pressure, local practices and prior experience (Levy and Williams, 2004; Gruppen and Frohna, 2002).

In fact, task-specific expertise has been shown to be a key variable in understanding differences in information processing – and thus task performance (Ericsson, 2006). There is ample research indicating that prolonged task experience helps novices develop into expert-like performers through the acquisition of an extensive, well-structured knowledge base as well as adaptations in cognitive processes to efficiently process large amounts of information in handling complex tasks. Research findings consistently indicate that these differences in cognitive structures and processes impact on proficiency and quality of task performance (Chi, 2006). For instance, a main characteristic of expert behaviour is the predominance of rapid, automatic pattern recognition in routine problems, enabling extremely fast and accurate problem solving (Klein, 1993; Coderre et al., 2003). When confronted with unfamiliar or complex problems, however, experts tend to take more time to gather, analyze and evaluate information in order to better understand the problem, whereas novices are more prone to start generating a problem solution or course of action after minimal information gathering (Ross et al., 2006; Voss et al., 1983). Another robust finding in expertise studies is that, compared with non-experts, experts see things differently and see different things. In general, experts make more inferences on information, clustering sets of information into meaningful patterns and abstractions (Chi et al., 1981; Feltovich et al., 2006). Studies on expert behaviour in medicine, for instance, show that experts have more coherent explanations for patient problems, make more inferences from the data and provide fewer literal interpretations of information (Van de Wiel et al., 2000). Similar findings were described in a study on teacher supervision (Kerrins and Cushing, 2000). Analysis of verbal protocols showed that inexperienced supervisors mostly provided literal descriptions of what they had seen on the videotape. More than novices, experienced supervisors interpreted their observations as well as made evaluative judgments, combining various information into meaningful patterns of classroom teaching. Overall, experts' observations focused on students and student learning, whereas non-experts focused more on discrete aspects of teaching.

Research findings also indicate that experts pay attention to cues and information that novices tend to ignore. For instance, experts typically pay more attention to contextual and situation-specific cues while monitoring and gathering information, whereas novices tend to focus on literal textbook aspects of a problem. In fact, automated processing by medical experts seems to heavily rely on contextual information (e.g. Hobus et al., 1987).

Finally, experts generally have better (more accurate) self-monitoring skills and greater cognitive control over aspects of performance where control is needed. Not only are experts able to devote cognitive capacity to self-monitoring during task performance, their richer mental models also enable them to better detect errors in their reasoning. Feltovich et al. (1984), for instance, investigated flexibility of experts versus non-experts on diagnostic tasks. Results showed that novices were more rigid and tended to adhere to initial hypotheses, whereas experts were able to discover that the initial diagnosis was incorrect and adjust their reasoning accordingly. In their study on expert-novice differences in teacher supervision, Kerrins and Cushing (2000) found that experts were more cautious in over- and underinterpreting what they were seeing. Although experts made more interpretative and evaluative comments, they more often qualified their comments with respect to both their interpretation of the evidence and the limitations of their task environment.

Based on the conceptual frameworks of cognition-based performance assessment and expertise research, it is perfectly conceivable that rater behaviour in WBA changes over time, due to increased task experience. Extrapolating findings from research in other domains, different levels of expertise may then be reflected in differences in task performance, which may have implications not only for utility of work-based assessments, but also for the way we select and train our raters. Given the increased significance of WBA in health professions education, the question can therefore be raised whether expertise effects as described also occur in performance assessment in the clinical domain. The present study aims to investigate cognitive processes related to judgment and decision making by raters observing performance in the clinical workplace. Verbal protocols; time spent on performance analysis and representation, and performance scores were analyzed to assess differences between experienced and non-experienced raters. More specifically, we explored 4 hypotheses that arose from the assumption that task experience determines information processing by raters. Firstly, we expected experienced raters to take less time, compared to non-experienced raters, in forming initial representations of trainee performance when observing prototypical behaviours, but more time when more complex behaviours are involved. Secondly, we expected experienced raters to pay more attention to situation-specific cues in the context of the rating task, such as patient or case specific cues; the setting of the patient encounter and ratee experience (phase of training). Thirdly, verbal protocols of experienced raters were expected to contain more inferences (interpretations) and fewer literal descriptions of behaviours. Finally, experienced raters were expected to generate more self-monitoring statements during performance assessment.

Method

Participants

The participants in our study were GP-supervisors who were actively involved as supervisor-assessor in general practice residency training. General practice training in the Netherlands has a long tradition of systematic direct observation and assessment of trainee performance throughout the training program. GP-supervisors are all experienced general practitioners, continuously involved in supervision of trainees on a day-to-day basis. They are trained in assessment of trainee performance.

In our study, we defined the level of expertise as the number of years of task-relevant experience as a supervisor-rater. Since there is no formal equivalent of elite rater performance we adopted a relative approach to expertise. This approach assumes that novices develop into experts through extensive task experience and training (Chi, 2006; Norman et al., 2006). In general, about 7 years of continuous experience in a particular domain is necessary to achieve expert performance (e.g. Arts et al., 2006). Registered GP-supervisors with different levels of supervision experience were invited to voluntarily participate in our study; a total of 34 GP-supervisors participated. GP-supervisors with at least 7 years of experience as supervisor-rater were defined as 'experts'. The 'expert group' consisted of 18 GP-supervisors (number of years of experience *M* = 13.4; *SD* = 5.9); the 'non-expert group' consisted of 16 GP-supervisors (number of years of experience *M* = 2.6; *SD* = 1.2). Levels of experience between both groups differed significantly ($t(32)=7.2$, $p < .001$). Participants received financial compensation for their participation.

Rating stimuli

The participants watched two DVDs, each showing a final-year medical student in a 'real-life' encounter with a patient. The DVDs were selected purposefully with respect to both patient problems and students' performance. Both DVDs presented 'straightforward' patient problems that are common in general practice: atopic eczema and angina pectoris. These cases were selected to ensure that all participants (both experienced and non-experienced raters) were familiar with required task performance. DVD 1 –atopic eczema- lasted about six minutes and presented a student showing prototypical and clearly substandard behaviour with respect to communication and interpersonal skills. This DVD was considered to present a non-complex rating task. DVD 2 –angina pectoris- lasted about eighteen minutes and was considered to present a complex rating task with the student showing more complex behaviours with respect to both communication and patient management. Permission had been obtained from the students and the patients to record the patient encounter and use the recording for research purposes.

Rating forms

The participants used two instruments to rate student performance (Figures 6.1 and 6.2): a one-dimensional, *overall* rating of student performance on a five-point Likert scale (1 = poor to 5 = outstanding) (R1), and a list of six clinical competencies (history taking; physical examination; clinical reasoning and diagnosis; patient management; communication with the patient; and professionalism), each to be rated on a five-point Likert scale (1=poor to 5=outstanding) (R2). Rating scales were kept simple to allow for maximum idiosyncratic cognitive processing. The participants were not familiar with the rating instruments and had not been trained in their use.

Research procedure and data collection

We followed standard procedures for verbal protocol analysis to capture cognitive performance (Chi, 1997)[1]. Before starting the first DVD, participants were informed about procedures and received a set of verbal instructions. Raters were specifically asked to "think aloud" and to verbalize all their thoughts as they emerged, as if they were alone in the room. If a participant were silent for more than a few seconds, the research assistant reminded him or her to continue.

Overall performance

- □ poor
- □ borderline
- □ satisfactory
- □ good
- □ outstanding

Figure 6.1 1-dimensional *overall* performance rating (R1)

History taking (*accurate, efficient*)

1	2	3	4	5	NA
poor	borderline	satisfactory	good	outstanding	

Physical Examination (*logical sequence, appropriate, informs patient*)

1	2	3	4	5	NA
poor	borderline	satisfactory	good	outstanding	

Clinical Reasoning / Diagnosis (*interpretation findings, judgment, efficiency*)

1	2	3	4	5	NA
poor	borderline	satisfactory	good	outstanding	

Patient Management (*adequate, addresses patient's needs/concerns)*

1	2	3	4	5	NA
poor	borderline	satisfactory	good	outstanding	

Communication with patient (*structure, communication skills, empathy*)

1	2	3	4	5	NA
poor	borderline	satisfactory	good	outstanding	

Professionalism (*organization, efficiency, respect, attends patient's needs)*

1	2	3	4	5	NA
poor	borderline	satisfactory	good	outstanding	

Figure 6.2 6-dimensional global rating scale clinical competencies (R2)

Permission to audiotape the session was obtained. For each of the DVDs the following procedure was used:

1. DVD starts. The participant signals when he or she feels able to judge the student's performance, and the time from the start of the DVD to this moment is recorded (T1). T1 represents the time needed for problem representation, i.e. initial representation of trainee performance.
2. The DVD is stopped at T1. The participant verbalizes his/her first judgment of the trainee's performance (verbal protocol (VP) 1).

3. The participant provides an overall rating of performance on the one-dimensional rating scale (R1T1), thinking aloud while filling in the rating form (VP2).
4. Viewing of the DVD is resumed from T1. When the DVD ends (T2), the participant verbalizes his/her judgment (VP3) and provides an overall rating (R1T2).
5. The participant fills in the multidimensional rating form (R2) for one of the DVDs (alternately DVD 1 or DVD 2) and verbalizes his or her thoughts while doing so (VP4).

We used a balanced design to control for order effects; the participants within each group were alternately assigned to one of two viewing conditions with a different order of the DVDs. All the audiotapes were transcribed verbatim.

Data analysis

The transcriptions of the verbal protocols were segmented into phrases by one of the researchers (MG). Segments were identified on the basis of semantic features (i.e. content features -as opposed to non content features such as syntax). Each segment represented a single thought, idea or statement (see Table 6.1 for some examples). Each segment was assigned to coding categories, using software for qualitative data analysis (Atlas.ti 5.2). Different coding schemes were used to specify 'the nature of the statement'; 'type of verbal protocol' and 'clinical presentation' (Table 6.1). The coding categories for 'nature of statement' were based on earlier studies in expert-novice information processing (Kerrins and Cushing, 2000; Boshuizen, 1989; Sabers et al., 1991) and included 'description', 'interpretation', 'evaluation', 'contextual cue' and 'self-monitoring'. Repetitions were coded as such.

All verbal protocols were coded by two independent coders (MG and research assistant). Inter-coder agreement based on five randomly selected protocols was only moderate (Cohen's kappa 0.67), and therefore the two coders coded all protocols independently and afterwards compared and discussed the results until full agreement on the coding was reached.

The data were exported from Atlas.ti to SPSS 17.0. For each participant, the numbers of statements per coding category were transformed to percentages in order to correct for between-subject variance in verbosity and elaboration of answers. Because of the small sample sizes and non-normally distributed data, non-parametric tests (Mann-Whitney U) were used to estimate the differences between the two groups in the time to initial representation of performance (T1); the nature of the statements, and performance ratings per DVD. We calculated effect sizes by using the formula $ES = Z/\sqrt{N}$ as is suggested for non-parametric comparison of two independent samples, where Z is the z-score of the Mann-Whitney statistic and N is the total sample size (Field, 2009, p.550). Effect sizes equal to 0.1, 0.3, and 0.5, respectively, indicate a small, medium, and large effect. For within-group differences of overall ratings (R1T1 versus R1T2) the Wilcoxon signed rank test was applied.

Table 6.1 *Verbal protocol coding schemes*

Nature of statement

1. Descriptions: (literal) descriptions of student behaviour (“he is smiling to the patient”; “he asks if this happened before”)
2. Inferences: interpretations and abstractions of performance (“he is an authoritarian doctor”; “he is clearly a young professional”; “it seems that he takes no pleasure in being a doctor”)
3. Evaluations: normative judgments, referring to implicit or explicit standards (“his physical examination skills are very poor”; “overall, his performance is satisfactory”)
4. Contextual cue: remarks referring to case-specific or context-specific cues such as patient characteristics, setting of the patient encounter, context of the assessment task (“this patient is very talkative”;“this looks like a hospital setting, not general practice”; “he is being videotaped”)
5. Self-monitoring: reflective remarks, nuancing (“although I am not sure if I saw this correctly”; “on hindsight I shouldn’t have…” “…….but on the other hand most senior residents do not know how to handle these problems either”); self-instruction and structuring of rating process (“first I am going to look at ….”; “when evaluating performance I always look at atmosphere and balance”); explication of standards and performance theory (“one should always start with open-ended questions”; “from a first-year resident I expect……”)
6. Residual category: repetitions, remarks not directly related to the rating task (e.g. statements related to the experiment; supervisory interventions)

Clinical presentation

1. Dermatological problem (DVD 1)
2. Cardiological problem (DVD 2)

Verbal protocol

VP1: verbal protocol at T1; initial representation of student behaviour

VP2: verbal protocol at T1, while filling out the one-dimensional rating scale (overall judgment)

VP3: verbal protocol at T2, after viewing DVD; overall judgment of student performance while filling out one-dimensional rating scale

VP4: verbal protocol while filling out 6-dimensional rating scale.

Results

Table 6.2 shows the results for the time to problem representation (T1) and the overall performance ratings for each DVD. Time to T1 was similar for experienced and non-experienced raters when observing prototypical behaviour (DVD 1). However, when observing the more complex behavioural pattern in DVD 2, experienced raters took significantly longer time for monitoring and gathering of information, whereas there was only minimal increase in time for non-experts (*U* = 79.00*, p* = .03, *ES* = 0.38).

Table 6.2 *Time needed for problem representation (T1) and performance ratings per DVD, for each group of raters*

	DVD 1 (prototypical, derma case)		DVD 2 (complex, cardio case)	
Variable	Experts (N = 18)	Non-experts (N = 16)	Experts (N = 18)	Non-experts (N = 16)
T1 (seconds)	112.0 (121)	109.5 (237)	260.0 (308)	139.0 (110)
R1T1 (rating at T1)	2.0 (2)^a	2.0 (2)	3.0 (1)^a	3.0 (1)^a
R1T2 (rating at T2, after viewing entire DVD)	2.0 (1)^b	2.0 (1)	2.0 (2)^b	2.5 (1)^b

Note Presented are the median and the inter-quartile range (in parentheses). Experts take significantly (*U* = 79.00, *p* = .03, *ES* = 0.38) more time for monitoring and gathering of information than novices when observing performance on DVD 2 (cardio case). Rating scores are based on a 5-point scale (1 = poor, 5 = outstanding). Values in the same column (DVD 1 and DVD 2 resp.) with different superscripts differ significantly (Wilcoxon Signed Ranks test, *p* < .05).

Table 6.2 shows non-significant differences between the two groups in the rating scores. A Wilcoxon signed ranks test, however, showed significant within-group differences between the rating scores at T1 and T2. In the expert group these differences were significant for both the dermatology case (*Z* = -2.31, *p* = .02, *ES* = 0.40) and the cardiology case (*Z* = -2.95, *p* = .003, *ES* = 0.51). In the non-expert group, significant differences were found for the cardiology case only (*Z* = -2.49, *p* = .01, *ES* = 0.43). The impact of the differences in rating scores at T1 resp. T2 is illustrated by (significant) shifts in the percentage of ratings representing a 'fail' (R1 ≤ 2). In the expert group, the proportion of failures for the dermatology case was 61% at T1 versus 89% at T2. For the cardiology case the proportion of failures shifted from 11% (T1) to 56% at T2 in the expert group, and from 6% (T1) to 50% at T2 in the non-expert group.

Table 6.3 presents the percentages (median, inter-quartile range) for the nature of the statements for each group, by verbal protocol and across all protocols (= overall, VP1+VP2+VP3+VP4). Overall, the experienced raters generated significantly more inferences or interpretations of student behaviour (*U* = 62.5, *p* = .005, *ES* = 0.48), whereas non-experts provided more descriptions (*U* = 68.5, *p* = .009, *ES* = 0.45). The verbal protocols after viewing the entire DVD (VP3) showed similar and significant differences between experienced and non-experienced raters with respect to interpretations (*U* = 71.5, *p* = .01, *ES* = 0.43) and descriptions of behaviours (*U* = 73, *p* = .01, *ES* = 0.42). Experienced raters also generated significantly more interpretations when filling out the six-dimensional global rating scale (*U* = 63, *p* = .004, *ES* = 0.48).

Table 6.3 also shows that experienced raters generated more references to context-specific and situation-specific cues. This difference was significant at T1 (*U* = 83, *p* = .04, *ES* = 0.37), and similar and near-significant (*U* = 89, *p* = .06) for the overall protocols and protocol VP3.

Evaluations showed no significant differences, except for VP2, with experienced raters generating significantly more evaluations (*U* = 87,5, *p* = .05, *ES* = 0.34).

No significant between-group differences were found with respect to self-monitoring.

Table 6.3 *Percentages of statements in verbal protocols for experienced raters (Exp) and non-experienced raters (Non-Exp)*

Variable	Overall		VP1		VP2		VP3		VP4	
	Exp	Non-Exp	Exp	Non-Exp	Exp	Non-Exp	Exp	Non-Exp	Exp	Non-Exp
	Mdn (IQR)	*Mdn (IQR)*	*Mdn (IQR)*	*Mdn (IQR)*	*Mdn (IQR)*	*Mdn (IQR)*	*Mdn (IQR)*	*Mdn (IQR)*	*Mdn (IQR)*	*Mdn (IQR)*
Descriptions	19.8 (13.2)	25.3 (11.0)[a]	19.8 (20.1)	26.1 (12.8)	10.8 (18.2)	10.7 (18.6)	16.9 (16.3)	29.4 (13.3)[a]	18.2 (16.5)	20.4 (10.6)
Inferences	19.0 (7.9)	14.7 (5.1)[a]	38.9 (22.9)	37.5 (21.2)	14.8 (25.6)	20.0 (25.1)	14.5 (9.8)	6.8 (9.2)[a]	13.5 (14.0)	5.6 (4.6)[a]
Evaluations	24.4 (10.4)	24.9 (4.7)	7.6 (16.0)	5.4 (10.1)	33.3 (15.1)	18.2 (20.0)[a]	24.9 (16.3)	25.9 (12.6)	35.9 (17.3)	41.3 (17.3)
Contextual cues	12.9 (7.7)	10.4 (7.3)	13.2 (7.7)	6.1 (14.6)[a]	8.1 (13.5)	.0 (13.7)	18.3 (15.0)	9.8 (11.1)	10.0 (9.1)	8.8 (8.6)
Self-monitoring	20.4 (8.8)	20.5 (10.7)	15.2 (10.8)	16.7 (10.6)	27.9 (22.7)	40.4 (35.9)	20.2 (13.2)	19.4 (10.4)	17.2 (14.6)	13.5 (13.4)

Note Presented are the median en inter-quartile range (in parentheses).

VP1 = verbal protocol at T1; initial representation of student behaviour; VP2 = verbal protocol at T1, while filling out the one-dimensional rating scale (overall judgment); VP3 = verbal protocol at T2, after viewing entire DVD, overall judgment of student performance while filling out one-dimensional rating scale; VP4 = verbal protocol while filling out 6-dimensional rating scale.

[a] indicates significant differences between experienced and non-experienced raters [Mann-Whitney U test, p < .05].

Discussion

Based on expertise research in other domains, we hypothesized that experienced raters would differ from non-experienced raters with respect to cognitive processes that are related to judgment and decision making in workplace-based assessment.

As for the differences in the time taken to arrive at the initial representation of trainee performance, the results partially confirm our hypothesis. It is contrary to our expectations that the expert raters took as much time as the non-expert raters with the case presenting prototypical behaviour, but our expectations are confirmed for the case with complex trainee behaviour, with the experts taking significantly more time than the non-experts. This finding is consistent with other findings on expertise research (Ericsson and Lehmann, 1996). Whereas non-experienced raters seem to focus on providing a correct solution (i.e. judgments or performance scores) irrespective of the complexity of the observed behaviour, expert raters take more time to monitor, gather and analyze the information before arriving at a decision on complex trainee performance. Our non-significant results with respect to prototypical behaviour may be explained by the rating stimulus in our study. The dermatology case may have been too short, and the succession of typical student behaviours too quick to elicit differences. Moreover, the clearly substandard performance in the stimulus may have elicited automatic information processing and pattern recognition in both groups (Eva, 2004). Our results for the cardiology case, however, confirm that, with more complex behaviours, experienced raters seem to differ from non-experienced raters with respect to their interpretation of initial information -causing them to search for additional information and prolonged monitoring of trainee behaviour.

As for the verbal protocols, the overall results appear to confirm the hypothesized differences between expert and non-expert raters in information processing while observing and judging performance. Compared to non-experienced raters, experienced raters generated more inferences on information and interpretations of student behaviours, whereas non-experienced raters provided more literal descriptions of the observed behaviour. These findings suggest that non-experienced raters pay more attention to specific and discrete aspects of performance, whereas experienced raters compile different pieces of information to create integrated chunks and meaningful patterns of information. Again, this is consistent with other findings from expertise research (Chi, 2006). Our results also suggest that expert raters have superior abilities to analyze and evaluate contextual and situation-specific cues. The raters in our study appeared to pay more attention to contextual information and to take a broader view, at least in their verbalizations of performance judgments. They integrate relevant background information and observed behaviours into comprehensive performance assessments. The differences between experts and non-experts were most marked at the initial stage of information gathering and assessment of performance (VP1). The setting of the patient encounter, patient characteristics and the context of the assessment task all seem to be taken into account in the experts' initial judgments.

These findings suggest that expert raters possess more elaborate and coherent mental models of performance and performance assessment in the clinical workplace. Similar expert-novice differences have been reported in other domains. Cardy et al. (1987), for instance, found that experienced raters in personnel management use more and more sophisticated categories for describing job performance. Our findings are in line with many other studies in expertise development, which consistently demonstrate that compared with novices, experts have more elaborate and well-structured mental models, replete with contextual information.

The results of our study showed that, within groups, the initial ratings at T1 differed significantly from the ratings after viewing the entire DVD (T2). Thus our findings suggest that both expert and non-expert raters continuously seek and use additional information, readjusting judgments while observing trainee performance. Moreover, this finding points to the possibility that rating scores, provided after brief observation, may not accurately reflect overall performance. This could have consequences for guidelines for minimal observation time and sampling of performance in WBA. Our results did not reveal significant differences in rating scores between experts and non-experts. We were therefore not able to confirm previous research findings in industrial psychology demonstrating that expert raters provide more accurate ratings of performance compared with non-experts (e.g. Lievens, 2001). Possible explanations are that, as a result of previous training and experience in general practice, both groups may have common notions of what constitutes substandard versus acceptable performance in general practice. Shared frames of reference, a rating scale that precludes large variations in performance scores and the small sample size may have caused the equivalent ratings in both groups.

Contrary to our expectations, the experts in our study do not appear to demonstrate more self-monitoring behaviour while assessing performance. An explanation might be that our experimental setting, in which participants were asked to think aloud while providing judgments about others, induced more self-explanations. The task of verbalizing thoughts while filling out a rating scale and providing a performance score may have introduced an aspect of accountability into the rating task, with both experienced and non-experienced raters feeling compelled to explain and justify their actions despite being instructed otherwise. These self-explanations and justifications of performance ratings may also explain the absence of any significant differences in rating scores between the groups. Several studies have shown that explaining improves subjects' performance (e.g. Chi et al., 1994). And research into performance appraisal in industrial organizations has demonstrated that raters who are being held accountable provide more accurate rating scores (e.g. Mero et al., 2003). The think aloud procedure may therefore have resulted in fairly accurate rating scores in both groups. This explanation is substantiated by the comments of several raters on effects of verbalization [e.g. "*If I had not been forced to think aloud, I would have given a 3 (satisfactory), but if I now reconsider what I said before, I want to give a 2 (borderline)*"].

What do our findings mean and what are the implications for WBA?

Our findings offer indications that in workplace-based assessment of clinical performance expertise effects occur that are similar to those reported in other domains, providing support for cognition-based models of assessment as proposed by Feldman (1981) and others.

There are several limitations to our study. Participants in our study were all volunteers and therefore may have been more motivated to carefully assess trainees' performance. Together with the experimental setting of our study, this may limit generalization of our findings to raters in 'real life' general practice. Real life settings are most often characterized by time constraints, conflicting tasks and varying rater commitment, which may all impact on rater information processing. Another limitation of our study is the small sample size, although the sample used is not uncommon in qualitative research of this type. Also, statistical significant differences emerge despite the relatively small sample size and the use of less powerful, but more robust non-parametric tests. Finally, we used only years of experience as a measure of expertise; other variables such as actual supervisor performance, commitment to teaching and assessment, or reflectiveness were not measured or controlled for. Time and experience are clearly important variables in acquiring expertise, though. The purpose of our study was not to identify and elicit superior performance of experts. Rather, we investigated whether task-specific experience affects the way in which raters process information when assessing performance. In this respect, our relative approach to expertise is very similar to approaches in expertise research in the domain of clinical reasoning in medicine (Norman, 2006).

If our findings reflect research findings from expertise studies in other domains, this may have important implications for WBA. Our study appears to confirm the existence of differences in raters' knowledge structures and reasoning processes resulting from training as well as personal experience. Such expert-novice differences may impact the feedback that is given to trainees in the assessment process.

Firstly, more enriched processing and better incorporation of contextual cues by experienced raters can result in qualitatively different, more holistic feedback to trainees, focusing on a variety of issues. Expert raters seem to take a broader view, interpreting trainee behaviour in the context of the assessment task and integrating different aspects of performance. This enables them to give meaning to what is happening in the patient encounter. Non-experienced raters on the other hand may focus more on discrete 'checklist' aspects of performance. Similar findings have been reported by Kerrins and Cushing (2000) in their study on supervision of teachers.

Secondly, thanks to more elaborate performance scripts, expert raters may rely more often on top-down information processing or pattern recognition when observing and judging performance -especially when time constraints and/or competing responsibilities play a role. As a consequence, expert judgments may be driven by general, holistic impressions of performance neglecting behavioural detail (Murphy and Balzer, 1986; Lievens, 2001), whereas non-experienced raters may be more accurate at the behavioural level. However, research in other domains has shown that, despite being likely to chunk information under normal conditions, experts do not lose their ability to use and recall 'basic' knowledge underlying reasoning and decision making (Schmidt and Boshuizen, 1993). Moreover, research findings indicate that experts demonstrate excellent recall of relevant data when asked to process a case deliberately and elaborately (Norman et al., 1989; Wimmers et al., 2005). Similarly, when *obliged* to process information elaborately and deliberately, experienced raters may be as good as non-experienced raters in their recall of specific behaviours and aspects of performance. Optimization of WBA may therefore require rating procedures and formats that force raters to elaborate on their judgments and substantiate their ratings with concrete and specific examples of observed behaviours.

Finally, our findings may have consequences for rater training, not only for novice raters, but for more experienced raters as well. Clearly, there is a limit to what formal training can achieve and rater expertise seems to develop through real world experience. Idiosyncratic performance schemata are bound to develop as a result of personal experiences, beliefs and attitudes. Development of shared mental models and becoming a *true* expert, however, may require deliberate practice with regular feedback and continuous reflection on strategies used in judging (complex) performance in different (ill-defined) contexts (Ericsson, 2004).

Further research should examine whether our findings can be reproduced in other settings. Important areas for study are the effects of rater expertise on feedback and rating accuracy. Is there a different role for junior and senior judges in WBA? Our findings also call for research into the relationship between features of the assessment system, such as rating scale formats, and rater performance. Rating scale formats affect cognitive processing in performance appraisal to the extent that the format is more or less in alignment with raters' "natural" cognitive processes. Assuming that raters' cognitive processes vary with experience, it is to be expected that different formats will generate differential effects on information processing in raters with different levels of experience. For instance, assessment procedures which focus on detailed and complete registration of ratee behaviours may disrupt the automatic, top-down processing of expert raters, resulting in inaccurate ratings. We need to understand which rating formats facilitate or hinder rating accuracy and provision of useful feedback at different levels of rater expertise. There is also increasing evidence that rater behaviour is influenced by factors like trust in the assessment system, rewards and threats (consequences of providing low or high ratings), organizational norms, values, etc. (Murphy and Cleveland, 1995). These contextual factors may lead to purposeful 'distortion' of ratings. Future research should therefore include field research to investigate possible effects of these contextual factors on decision making. Finally we wish to emphasize the need for more in-depth and qualitative analysis of raters' reasoning processes in performance assessment. How do performance schemata of experienced raters differ from those of non-experienced raters? How do raters combine and weigh different pieces of information when judging performance? How are performance schemata and theories linked to personal beliefs and attitudes?

In devising measures to optimize WBA we should first and foremost take into account that raters are not interchangeable measurement instruments, as is generally assumed in the psychometric assessment framework. In fact, a built-in characteristic of cognitive approaches to performance assessment is that raters' information processing is guided by their 'mental models' of performance and performance assessment. Our study shows that raters' judgment and decision making processes change over time due to task experience, supporting the need for research as described above.

Notes

[1] Verbal protocols refer to the collection of participants' verbalizations of their thoughts and behaviours, during or immediately after performance of cognitive tasks. Typically, participants are asked to "think aloud" and to verbalize all their thoughts as they emerge, without trying to explain or analyze those thoughts (Ericsson and Simon, 1993). Verbal analysis is a methodology for quantifying the subjective or qualitative coding of the contents of these verbal utterances (Chi, 1997). Chi (1997) describes the specific technique for analyzing verbal data as consisting of several steps, excluding collection and transcription of verbal protocols. These steps, as followed in our research, are: defining the content of the protocols; segmentation of protocols; development of a coding scheme; coding the data and refining coding scheme if needed; resolving ambiguities of interpretation; and analysis of coding patterns.

References

Arts, J.A.R.M., Gijselaers, W.H. & Boshuizen, H.P.A. (2006). Understanding managerial problem-solving, knowledge use and information processing: Investigating stages from school to the workplace. *Contemporary Educational Psychology, 31*(4), 387-410.

Barneveld van, C. (2005). The dependability of medical students' performance ratings as documented on in-training evaluations. *Academic Medicine*, *80*(3), 309-312.

Boshuizen, H.P.A. (1989). *The development of medical expertise; a cognitive-psychological approach.* Doctoral dissertation. Maastricht: Rijksuniversiteit Limburg.

Cardy, R.L., Bernardin, H.J., Abbott, J.G., Senderak, M.P. & Taylor K. (1987). The effects of individual performance schemata and dimension familiarization on rating accuracy. *Journal of Occupational Psychology*, *60*, 197–205.

Chi, M. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences, 6*(3), 271-315.

Chi, M.T.H. (2006). Two approaches to the study of experts' characteristics. In: K.A. Ericsson, N. Charness, P.J. Feltovich & R.R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 21-30). Cambridge: Cambridge University Press.

Chi, M.T.H., de Leeuw, N., Chiu, M.H. & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *5*, 121-152.

Chi, M.T.H., Feltovich, P.J. & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science: A multidisciplinary Journal, 5*(2), 121-52.

Coderre, S., Mandin, H., Harasym, P.H. & Fick, G.H. (2003). Diagnostic reasoning strategies and diagnostic success. *Medical Education*, *37*, 695-703.

Cunnington, J. & Southgate, L. (2002). Relicensure, recertification and practice-based assessment. In: G.R. Norman, C.P.M. van der Vleuten & D.I. Newble (Eds.), *International Handbook of Research in Medical Education* (pp. 883-912). Dordrecht: Kluwer Academic Publishers.

DeNisi, A.S. (1996). *Cognitive Approach to Performance Appraisal: A Program of Research*. New York: Routledge.

Ericsson, K.A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine*, *79*(10), S70-81.

Ericsson, K.A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In: K.A. Ericsson, N. Charness, P.J. Feltovich & R.R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 683-704). Cambridge: Cambridge University Press.

Ericsson, K.A. & Lehmann, A.C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annu. Rev. Psychol*., *47,* 273–305.

Ericsson, K.A. & Simon, H.A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Eva, K.W. (2004). What every teacher needs to know about clinical reasoning. *Medical Education*, *39*, 98-106.

Feldman, J.M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, *66*(2), 127-148.

Feltovich, P.J., Johnson, P.E., Moller, J.H. & Swanson, D.B. (1984). LCS: The role and development of medical knowledge in diagnostic expertise. In: W.J. Clancey & E.H. Shortliffe (Eds.), *Readings in medical artificial intelligence: The first decade* (pp.275-319). Reading, MA: Addison Wesley.

Feltovich, P.J., Prietula, M.J. & Ericsson, K.A. (2006). Studies of expertise from psychological perspectives. In: K.A. Ericsson, N. Charness, P.J. Feltovich & R.R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 41-68). Cambridge: Cambridge University Press.

Field, A. (2009). *Discovering statistics using SPSS*. London, etc.: Sage Publications Ltd.

Gray, J.D. (1996). Global rating scales in residency education. *Academic Medicine, 71*, S55-61.

Gruppen, L.D. & Frohna, A.Z. (2002). Clinical reasoning. In: G.R. Norman, C.P.M. van der Vleuten & D.I. Newble DI (Eds.), *International Handbook of Research in Medical Education* (pp. 205-30). Dordrecht: Kluwer Academic Publishers.

Hawe, E. (2003). It's pretty difficult to fail: The reluctance of lecturers to award a failing grade. *Assessment and Evaluation in Higher Education, 28*(4), 371–382.

Hobus, P.P., Schmidt, H.G., Boshuizen, H.P. & Patel, V.L. (1987). Contextual factors in the activation of first diagnosis hypotheses: Expert-novice differences. *Medical Education, 21*, 471-76.

Kerrins, J.A. & Cushing, K.S. (2000). Taking a second look: Expert and novice differences when observing the same classroom teaching segment a second time. *Journal of Personnel Evaluation in Education*, *14*(1), 5-24.

Klein, G.A. (1993). A recognition primed decision (RPD) model of rapid decision making. In: G.A. Klein, J. Orasanu, R. Calderwood & C.E. Zsambok (Eds.), *Decision-making in action: Models and Methods* (pp.138-147). Norwood, NJ: Ablex.

Kreiter, C.D. & Ferguson, K.J. (2001). Examining the generalizability of ratings across clerkships using a clinical evaluation form. *Evaluation & The Health Professions, 24*, 36-46.

Levy, P.E. & Williams, J.R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management*, *30*, 881-905.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, *86*(2), 255-264.

Mero, N.P., Motowidlo, S.J. & Anna, A.L. (2003). Effects of accountability on rating behavior and rater accuracy. *Journal of Applied Social Psychology*, *33*(12), 2493-2514.

Murphy, K.R. & Balzer, W.K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluation: Consequences for rating accuracy. *Journal of Applied Psychology*, *71*, 39–44.

Murphy, K.R. & Cleveland, J.N. (1995). *Understanding Performance Appraisal: Social, Organizational and Goal-based Perspectives*. Thousand Oaks, CA: Sage Publications.

Norcini, J.J. (2005). Current perspectives in assessment: The assessment of performance at work. *Medical Education, 39*, 880-89.

Norman, G., Eva, K., Brooks, L. & Hamstra, S. (2006). Expertise in medicine and surgery. In: K.A. Ericsson, N. Charness, P.J. Feltovich & R.R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 339-354). Cambridge: Cambridge University Press.

Norman, G.R., Brooks, L.R. & Allen, S.W. (1989). Recall by expert medical practitioners and novices as a record of processing attention. *J Exp Psychol Learn Mem Cogn*, *15*, 1166-74.

Ross, K.G., Shafer, J.L. & Klein, G. (2006). Professional Judgments and "Naturalistic Decision Making". In: K.A. Ericsson, N. Charness, P.J. Feltovich & R.R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 403-420). Cambridge: Cambridge University Press.

Sabers, D.S., Cushing, K.S. & Berliner, D.C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensionality and immediacy. *American Educational Research Journal, 28,* 63-88.

Schmidt, H.G. & Boshuizen, H.P.A. (1993). On the origin of intermediate effects in clinical case recall. *Memory and Cognition*, *21*, 338-351.

Silber, C.G., Nasca, T.J., Paskin, D.L., Eiger, G., Robeson, M. & Veloski, J.J. (2004). Do global rating forms enable program directors to assess the ACGME competencies? *Academic Medicine, 79*, 549-556.

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

Voss, J.F., Tyler, S.W. & Yengo, L.A. (1983). Individual differences in the solving of social science problems. In: R. Dillon R & R. Schmeck (Eds.), *Individual differences in cognition* (pp. 205-232). New York: Academic Press.

Walsh, J.P. (1995). Managerial and organizational cognition: Notes from a trip down to memory lane. *Organizational Science*, *6*(3), 280-321.

Wiel van de, M.W.J., Boshuizen, H.P.A. & Schmidt, H.G. (2000). Knowledge restructuring in expertise development: Evidence from pathophysiological representations of clinical cases by students and physicians. *European Journal of Cognitive Psychology, 12*(3), 323-356.

Williams, R.G. & Dunnington, G.L. (2004). Prognostic value of resident clinical performance ratings. *J Am Coll Surg*, *199*, 620-27.

Williams, R.G., Klamen, D.A. & McCaghie, W.C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine, 15*(4), 270-292.

Wimmers, P.F., Schmidt, H.G., Verkoeijen, P.P.J.L. & Van de Wiel, M.W.J. (2005). Inducing expertise effects in clinical case recall. *Medical Education*, *39*(9), 949-57.

CHAPTER 7

Raters' performance theories and constructs in workplace-based assessment

Submitted for publication

Marjan J.B. Govaerts
Margje W.J. van de Wiel
Lambert W.T. Schuwirth
Cees P.M. van der Vleuten
Arno M.M. Muijtjens

Abstract

Objectives • Weaknesses in the nature of rater judgments are generally considered to compromise the utility of work-based assessment (WBA). In order to gain insight into the underpinnings of rater behaviours, we investigated how raters form impressions and make judgments about trainee performance. Using theoretical frameworks of social cognition and person perception, we explored raters' implicit performance theories ('role schemas'), use of task-specific performance schemas and the formation of person schemas during WBA. We furthermore explored effects of rater experience on schema-based processing.

Method • Experienced (N = 18) and non-experienced (N = 16) raters (GP-supervisors) watched two videotapes, each presenting a trainee in a 'real-life' patient encounter. Cognitive performance of raters was captured through think-aloud procedures and verbal protocol analysis. Qualitative data analysis was used to explore schema content and usage. We quantitatively assessed rater idiosyncrasy in the use of performance dimensions in WBA and we investigated effects of rater expertise on the use of (task-specific) performance schemas.

Results • Raters used different schemas in judging trainee performance. We developed a normative performance theory comprising seventeen inter-related performance dimensions. Levels of rater idiosyncrasy were substantial and unrelated to rater expertise. Significant differences between experienced and non-experienced raters were found with respect to use of task-specific performance schemas, suggesting that experienced raters have more differentiated performance schemas. Most raters started to develop person schemas the moment they began to observe trainee performance.

Conclusions • The findings further our understanding of processes underpinning judgment and decision making in WBA. Raters make and justify judgments based on personal theories and performance constructs. Raters' information processing seems to be affected by differences in rater expertise. The results of this study can help to improve rater training, the design of assessment instruments and decision making in WBA.

Introduction

Observation and assessment of trainees' performance in 'real-life' professional settings has been a cornerstone of health professions education for centuries. It is the potentially best way of collecting data and providing feedback on what trainees actually *do* in day-to-day practice. In fact, current assessment practices are characterized by a growing emphasis on workplace-based assessment (WBA), stimulated by widespread implementation of competency-based curricula; increasing demands for physician accountability and concerns about health care quality as well as calls for improved supervision and assessment of medical trainees (Davies, 2005; Norcini, 2005; Kogan et al., 2009; Holmboe et al., 2010).

Although there is general agreement that WBA is useful for formative assessment, its usefulness for summative assessment purposes is not undisputed (Norcini and Burch, 2007; McGaghie et al., 2009). Major concerns about the utility of WBA relate to its inherent subjectivity and the resulting weaknesses in the quality of measurement. In general, the idiosyncratic nature of (untrained) raters' judgments results in large differences between performance ratings, low inter- and intra-rater reliabilities and questionable validity of WBA (Albanese, 2000; Williams et al., 2003). More to the point, research into performance appraisals in various domains suggests that idiosyncratic rater effects account for substantial variance in performance ratings, ranging from 29% to over 50% (Viswesvaran et al., 1996; Scullen et al., 2000; Hoffmann et al., 2010). Consequently, attempts to improve work-based assessment tend to focus on minimizing the 'subjectivity factor' through standardization of assessment procedures and rater training. However, such measures have met with mixed success at best (Williams et al., 2003; Lurie et al., 2009; Holmboe et al., 2010; Green and Holmboe, 2010).

Research findings suggest that many reasons may underlie persistence in rater behaviour despite training and / or the use of worked out (detailed) assessment tools. Research in industrial and organizational psychology indicates that rating outcomes are likely determined by a complex and interrelated set of factors, such as affect and motivation, assessment purposes, time pressure, local practices (norms and values) as well as actual ratee performance (Murphy and Cleveland, 1995; Levy and Williams, 2004). Research findings also indicate that raters often seem to hold implicit performance theories, which may diverge from those specified by the organization (Borman, 1987; Ostroff and Ilgen, 1992; Uggerslev and Sulsky, 2008). Recent research by Ginsburg et al. (2010) suggests that in the medical domain assessment tools and theoretical models of professional competence may not adequately reflect supervisors' theories of work performance, resulting in 'blurring' of competency domains and seemingly invalid or inaccurate ("less authentic") performance ratings. In other words, there may very well be discrepancies between how we feel that raters *should* think or act (theory espoused), and what they actually think and *do* in practice (theory in use). Similarly, Holmboe et al. (2010) state that in fact "….*we know very little about effective faculty observation skills and behaviours*."
Ginsburg et al. conclude that it may make more sense to start by investigating what raters actually observe, experience and can comment on. In order to effectively improve WBA, we

clearly need a better understanding of the underpinnings of rater behaviour in the context of WBA.

Raters as social perceivers

It is clearly inherent in work-based assessments that all information must ultimately pass the cognitive filter represented by the rater (Landy and Farr, 1980; Smith and Collins, 2009). This implies that understanding evaluation of performance in real life is basically about understanding how raters form impressions and make inferences (e.g., judgments and decisions) about other people in interpersonal and social environments. There is, in fact, growing agreement that raters are to be seen as 'social perceivers', who provide 'motivated social judgments' when evaluating performance (Murphy and Cleveland, 1995; Klimosky and Donahue, 2001; Levy and Williams, 2004). A central assumption in this approach is that raters are active information processors who, within a dynamic and complex social setting, are challenged by the cognitive tasks of gathering information, interpretating and integrating information, and retrieval of information for judgment and decision making (DeNisi, 1996; Klimoski and Donahue, 2001; McCaghie et al., 2009). Raters' information processing is influenced by their understanding of (in)effective performance, personal goals, interactions with the ratee and others, as well as other factors in the social context of the assessment process (Uggerslev and Sulsky, 2008; Murphy et al., 2004; Govaerts et al., 2007). This view of how raters perceive and judge performance can be cast in theoretical frameworks of social perception, as an element of social cognition. In fact, performance assessment might be seen as a 'specific application of social perception' for specific purposes and much of raters' behaviours can be considered to be rooted in social perception phenomena (Klimoski and Donahue, 2001; Barness-Farrell, 2001).

Performance assessment and social perception

Findings from social perception research consistently indicate that when forming impressions and making judgments of others, social perceivers tend to use pre-existing knowledge structures, or so-called schemas. Schemas can be thought of as adaptive mechanisms that enable people to efficiently process information, especially in situations where information is incomplete, ambiguous or where there are situational constraints (e.g. time pressure, conflicting tasks). The schemas that most people use in social perception are the *role, event* and *person* schemas (Pennington, 2000, pp.69-75). A role schema can be defined as the sets of behaviours expected of a person in *a certain social position* (e.g. policeman, teacher, family physician). Event schemas describe what we normally expect from other people's behaviours *in specific social situations,* related to the predicted sequence of events in such a situation (e.g. a job interview or performance appraisal interview). Person schemas, in conclusion, reflect the inferences that we make about someone on the basis of (limited) available information, as we come to know them through verbal and non-verbal cues in their behaviour. Person schemas may include expected patterns of behaviour, personality traits and other inferences, such as conclusions about someone's knowledge base or social category (for instance, 'excellent performer' or 'poor performer'). These schemas together guide and influence the focus of our attention when we observe others, what we remember and how we use information in forming impressions and making judgments. The three types of schema should not be regarded as entirely distinct or separate: schemas are used interactively when we try to understand other people's behaviours (Pennington, 2000).

Key features of the framework we have described can easily be translated to the context of work-based performance assessment.

First, the literature (e.g. Borman, 1987; Ostroff and Ilgen, 1992; Uggerslev and Sulsky, 2008; Ginsburg et al., 2010) suggests that raters in work settings develop personal constructs or 'theories' of effective job performance in general. These 'performance theories' are very similar to role schemas in that they include sets or clusters of effective behaviours in relation to any number of performance dimensions considered relevant for the job. Since performance theories develop through (professional) experience, socialization as well as training, the content of performance theories is likely to vary between raters, resulting in varying levels of rater idiosyncrasy (Uggerslev and Sulsky, 2008).

Second, research findings indicate that the particular set of behaviours that is related to effective performance may differ from one task to another, depending on the setting and specific features of the task (e.g. Veldhuijzen et al., 2007). Veldhuijzen et al.'s study, for instance, clearly showed that physicians use different communication strategies depending on situational demands. It is therefore to be expected that raters, through prolonged job experience, develop highly differentiated performance schemas, each representing different sets of effective behaviours for various and differentiated job-related tasks and task settings. When raters are observing others during task performance, task- or situation-specific cues may trigger the use of task- or 'event'-specific schemas to judge performance, especially in more experienced raters.

Finally, when observing performance for assessment purposes, raters will inevitably develop 'person schemas' to organize their knowledge about individual ratees. Raters will interpret observations, integrate information, and make inferences, for instance about a ratee's knowledge base, level of competence, or behavioural disposition.

Raters are likely to use all three types of schema interactively when making judgments and decisions about performance by others: a rater's personal performance theory ('role schema'); normative expectations of task-specific behaviours (task-specific schema) as well as inferences about the ratee (person schema) may all influence assessment outcomes (Cardy et al., 1987; Borman et al., 1987).

The present study

Given the increased significance of WBA in health professions education, the question may be raised whether theoretical frameworks of social perception can be used to further our understanding of processes underlying judgment and decision making in performance assessments so as to improve the utility of assessment outcomes.

Building on the social perception framework as described above, the present study explored the use of schemas by physician-raters when assessing trainee performance in patient encounters. Qualitatively, verbal protocol analysis was used to explore:

- Raters' implicit role schemas (which we will refer to as 'performance theories');
- Raters' use of task-specific performance schemas; and
- Raters' formation of person schemas during observation and assessment of performance.

We quantitatively analyzed differences between raters in the performance dimensions used in judgment and decision making (rater idiosyncrasy) and how differences in rating experience affected schema-based processing. Specifically, we expected experienced raters to use more task-specific performance schemas compared to non-experienced raters.

Method

Participants

The participants in our study were GP-supervisors who were actively involved as supervisor-assessor in general practice residency training. In the Netherlands, postgraduate training in general practice has a long tradition of systematic direct observation and assessment of trainee performance throughout the training program. All GP-supervisors are experienced general practitioners, who continuously supervise trainees on a day-to-day basis. They are trained in assessment of trainee performance.

Registered GP-supervisors with different levels of supervision experience were invited to participate in our study. A total of 34 GP-supervisors participated. In line with findings from expertise research (e.g. Arts et al., 2006), GP-supervisors with at least seven years of experience as supervisor-rater were defined as 'experts'. The 'expert group' consisted of eighteen GP-supervisors (experience as GP: *M* = 26.3 years; *SD* = 5.0 years; supervision experience: *M* = 13.4 years; *SD* = 5.9 years); the 'non-expert group' consisted of sixteen GP-supervisors (experience as GP: *M* = 12.9 years; *SD* = 5.0 years; supervision experience: *M* = 2.6 years; *SD* = 1.2 years). Participants received financial compensation for their participation.

Research procedure and data collection

Participants watched two video cases (VCs), each showing a final-year medical student in a 'real-life' encounter with a patient. Participants had never met the medical students before. The VCs were selected purposively with respect to both patient problems and students' performance. Both VCs presented 'straightforward' patient problems that are common in general practice: atopic eczema and angina pectoris. These cases were selected to ensure that all participants (both experienced and non-experienced raters) were familiar with task-specific performance requirements. VC1 –atopic eczema- lasted about six minutes and presented a student showing prototypical and clearly substandard behaviour with respect to communication and interpersonal skills. VC2 – angina pectoris- lasted about eighteen minutes and presented a student showing complex, i.e. more differentiated behaviours with respect to both communication and patient management. Permission had been obtained from the students and the patients to record the patient encounter and use the recording for research purposes.

Cognitive performance of participants was captured through verbal protocol analysis (Chi, 1997). Before starting the first video, participants were informed about research procedures and received a set of verbal instructions. Raters were specifically asked to "think aloud" and to verbalize all their thoughts as they emerged, as if they were alone in the room. If a participant were silent for more than a few seconds, the research assistant reminded him or her to continue. Permission to audiotape the session was obtained.

For each of the VCs the following procedure was used:

1. Video is started. The participant signals when he or she feels able to judge the student's performance; the video is then stopped (T1). The participant verbalizes his/her first judgment of the trainee's performance (verbal protocol (VP)1).
2. Subsequently, the participant provides an overall rating of performance on a one-dimensional rating scale, thinking aloud while filling in the rating form (VP2).
3. Viewing of the video is resumed from T1. When the video ends (T2), the participant verbalizes his/her judgment (VP3) while providing a final overall rating.

We used a balanced design to control for order effects; the participants within each group were alternately assigned to one of two viewing conditions with a different order of the VCs.

Data analysis

All the audiotapes were transcribed verbatim. Verbal data were analyzed qualitatively (to explore raters' schemas in assessment of performance) and quantitatively (to assess differences in use of schemas between raters and rater groups).

Qualitative analysis • Firstly, we undertook open, bottom-up coding of all verbal protocols (VP1, VP2, and VP3 pooled) to explore the raters' performance theories and task-specific performance schemas. Two researchers with different professional backgrounds (MG; MD, medical educator and MvdW cognitive psychologist) began by coding transcripts independently, using an open thematic type of analysis to determine the dimensions of performance that raters actually used when judging trainee performance. Researchers met repeatedly to compare and discuss emergent coding structures, until the coding framework was stable. The final coding framework, which was considered to represent raters' aggregate performance theory (i.e. the set of dimensions used by the raters to evaluate performance) and the coding structures reflecting the use of task-specific schemas, was discussed with an experienced general practitioner in order to assess confirmability. This discussion did not result in any further changes of the coding structure.

To explore the use of person schemas, we used top-down, *a priori* coding. The coding categories for 'person schemas' were based on the theoretical framework as proposed by Klimoski and Donahue (2001), which describes five types of inference processes that are most common for judgment tasks: inferences regarding knowledge; traits; disposition (probable patterns of behaviour); intentions (immediate goals) and social category membership. To this framework, we added a separate category indicating the use of 'training phase' as a frame of reference in making judgements.

Table 7.1 presents the final coding framework, which was applied to all verbal protocols using software for qualitative data analysis (Atlas-ti 6.1).

Table 7.1 *Verbal protocol coding structures*

Performance theory: Performance dimensions and sub dimensions

1. Think and act like a general practitioner
2. Doctor-patient relationship
   - 2.1. Atmosphere
   - 2.2. Balanced patient-centeredness
     - 2.2.1. Develop and establish rapport
     - 2.2.2. Demonstrate appropriate confidence
     - 2.2.3. Demonstrate empathy / empathic behaviour, appropriate for problem
     - 2.2.4. Open approach
     - 2.2.5. Facilitating shared mind 1 = identifying reasons for consultation; exploring patient's perspective
     - 2.2.6. Facilitating shared mind 2 = explain rationale for questions, examinations; explain process; share own thinking
     - 2.2.7. Facilitating shared mind 3 = collaborative decision making
3. (Bio)medical aspects (disease)
   - 3.1. History taking
   - 3.2. Physical examination
   - 3.3. Diagnosis / differential diagnosis
   - 3.4. Patient management plan
4. Structuring of the consultation and time management

Task- (event-)specific schema

1. Identification of case-specific cues
   - 1.1. Specific aspects of the patient's problem / clinical presentation (e.g. this type of eczema poses very serious social problems to the patient)
   - 1.2. Specific aspects of the patient's behaviours (verbal as well as non-verbal; e.g. this patient is very talkative)
   - 1.3. Setting / context of the medical consultation (GP's office versus outpatient clinic)
2. Trainee behaviours (effective or ineffective) within performance domain X, explicitly related to case-specific cues
3. Effects of trainee behaviour on patient behaviour / doctor-patient relationship (positive or negative)

Person schema

1. Inferences regarding
   - 1.1. Knowledge base
   - 1.2. Personality traits (e.g. he is a very nice guy)
   - 1.3. Disposition (e.g. this trainee has a clinical method of working; finds it difficult to just lean back and listen to what patients are saying)
   - 1.4. Intention (e.g. he seems to be focused on the biomedical aspect of the patient's problem)
   - 1.5. Category (e.g. he is an authoritarian doctor; he will become an excellent doctor)
2. Phase of training (frame of reference for making judgments)

Quantitative analysis • In order to explore differences between raters with respect to the use of performance theories and task-specific performance schemas, the verbal protocols were reanalyzed using the coding framework as presented in Table 7.1. For this analysis, VP1 and VP2 were merged to create a single verbal protocol containing all verbal utterances at T1. The transcripts of the verbal protocols were segmented into phrases by one of the researchers (MG). Each segment represented a single coherent thought or statement about the trainee or trainee performance (e.g. description of a particular behaviour within a performance dimension, or a judgmental remark about overall effectiveness on a particular performance dimension). Statements about trainee performance were also coded along the dimension positive versus negative (effective versus ineffective behaviour). Repetitions were coded as such. Six randomly selected protocols were coded by two independent coders (MG, MvdW). Since inter-coder agreement proved to be high (> 90%), other protocols were coded by only one researcher (MG). The data were then exported from Atlas.ti to SPSS 17.0. In order to explore rater idiosyncrasy with respect to the use of performance theory, we calculated for each performance dimension the percentage of raters using that performance dimension. Percentages were calculated for each VC at T1 and T2. Levels of rater idiosyncrasy in relation to any performance dimension can be inferred from the percentage of raters using that dimension, with 0% and 100% indicating maximum inter-rater agreement, i.e. complete absence of idiosyncrasy, and 50% indicating maximum disagreement, i.e. maximum level of idiosyncrasy. So, the closer the percentage moves to 50%, the higher the level of idiosyncrasy. Additionally, the number of statements representing dimension-related performance (effective versus ineffective behaviours), was calculated for each of the performance dimensions.

Between-group differences with respect to the use of task-specific schemas were estimated by transforming the number of statements per coding category per rater to percentages to correct for between-subject variance in verbosity and elaboration of answers. Because of the small sample sizes and non-normally distributed data, non-parametric tests (Mann-Whitney *U*) were used to estimate differences between the two groups. We calculated effect sizes by using the formula $ES = Z/\sqrt{N}$ as is suggested for non-parametric comparison of two independent samples, where *Z* is the *z*-score of the Mann-Whitney statistic and *N* is the total sample size (Field, 2009, p.550). Effect sizes of 0.1, 0.3, and 0.5, indicate a small, medium, and large effect, respectively.

Results

We will first present results from qualitative data analysis, followed by results from quantitative analyses.

Performance theory

Analysis of verbal protocols resulted in the identification of seventeen performance dimensions, used by GP- raters when assessing trainee behaviour during patient encounters. GP-raters distinguished four main dimensions (representing ‘overall job effectiveness’, ‘doctor-patient relationship’, ‘handling of (bio)medical aspects’, and ‘structuring / time management’), and various sub-dimensions. Within the dimension ‘doctor-patient relationship’, two larger sub-dimensions could be identified. One sub-dimension included the sets of behaviours related to “creating a good atmosphere” for effective and efficient patient-doctor communication. This sub-dimension was considered by the raters at the beginning of the consultation in particular. The second sub-dimension (“balanced patient centeredness”) contained the sets of behaviours facilitating patient involvement throughout the consultation while at the same time ensuring that the physician, as a professional medical expert, remains in charge of the consultation.

The performance dimensions, their interrelationships and examples of performance-related behaviours are presented in Figure 7.1.

Although participants clearly distinguished between different dimensions, data analysis also showed that participants used dimensions interactively when judging performance effectiveness. For instance, when evaluating the doctor-patient relationship, raters also considered whether the trainee organized and planned the consultation adequately:

> "*In the beginning, he did very well to let the patient tell his story, but it took too long; he should have guided the patient in the right direction a bit sooner–although it is not bad at all to just sit and listen to what the patient has to say." (PP 25)*

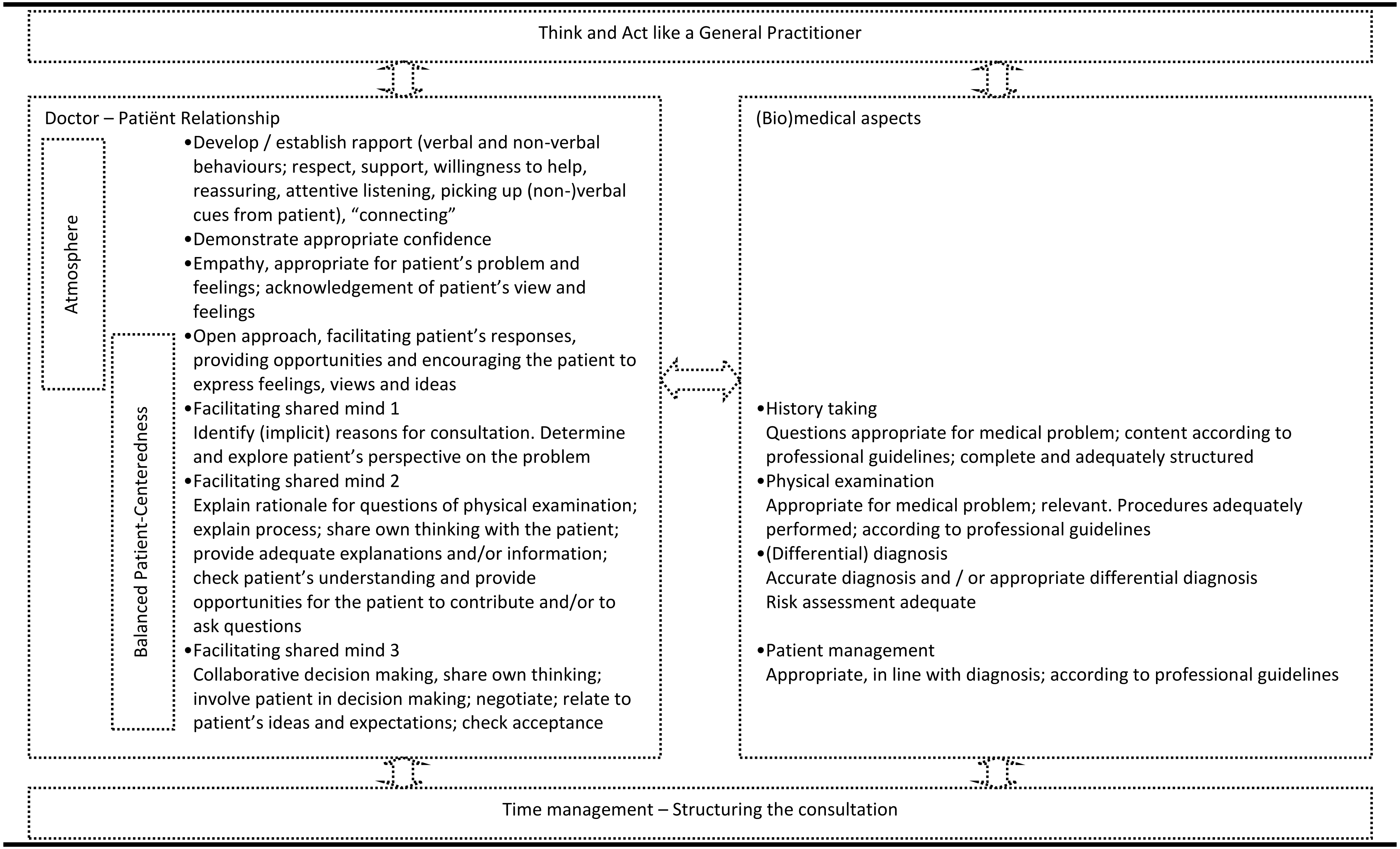
Similarly, when judging performance during physical examination, participants not only paid attention to technical skills and smoothness in performing the examination, but also took into account if and how the trainee communicated with the patient before and during the examination, as well as possible effects of the communication on the doctor-patient relationship:

> *"Physical examination is also not so good .. takes blood pressure, palpates abdomen, auscultation of abdomen,.. okay,…..but he examines the patient in dead silence. He doesn't tell the patient what he is doing nor what his findings are. This is not the way to gain the patient's trust." (PP 24)*

> *"Technically, physical examination seems to be adequate, but there was complete silence, no contact with the patient whatever. …..This might be another method to build trust with the patient, and it is missing completely." (PP3)*

Also in judgments of history taking or patient management, behaviours within the dimension "doctor-patient relationship" were considered as important as 'content-related' behaviours - within the dimension "(bio)medical aspects".

> *"His knowledge base seems to be adequate. From a cardiological perspective prescriptions are correct, but there is no link at all with the patient's view or feelings. …. I therefore doubt whether he is able to think like a general practitioner." (PP12)*



Think and Act like a General Practitioner

Doctor – Patiënt Relationship

Atmosphere

Balanced Patient-Centeredness

- Develop / establish rapport (verbal and non-verbal behaviours; respect, support, willingness to help, reassuring, attentive listening, picking up (non-)verbal cues from patient), "connecting"
- Demonstrate appropriate confidence
- Empathy, appropriate for patient's problem and feelings; acknowledgement of patient's view and feelings
- Open approach, facilitating patient's responses, providing opportunities and encouraging the patient to express feelings, views and ideas
- Facilitating shared mind 1
  Identify (implicit) reasons for consultation. Determine and explore patient's perspective on the problem
- Facilitating shared mind 2
  Explain rationale for questions of physical examination; explain process; share own thinking with the patient; provide adequate explanations and/or information; check patient's understanding and provide opportunities for the patient to contribute and/or to ask questions
- Facilitating shared mind 3
  Collaborative decision making, share own thinking; involve patient in decision making; negotiate; relate to patient's ideas and expectations; check acceptance

(Bio)medical aspects

- History taking
  Questions appropriate for medical problem; content according to professional guidelines; complete and adequately structured
- Physical examination
  Appropriate for medical problem; relevant. Procedures adequately performed; according to professional guidelines
- (Differential) diagnosis
  Accurate diagnosis and / or appropriate differential diagnosis
  Risk assessment adequate

- Patient management
  Appropriate, in line with diagnosis; according to professional guidelines

Time management – Structuring the consultation

Figure 7.1 Aggregate performance theory GP patient encounter

Table 7.2 *Person schemas*

	Derma case (VC1)				Cardiocase (VC2)			
	T1		T2		T1		T2	
	Exp (N = 16)	N-exp (N = 12)	Exp (N = 18)	N-exp (N = 16)	Exp (N = 18)	N-exp (N = 16)	Exp (N = 18)	N-exp (N = 16)
Percentage of raters making inferences (1.1-1.5)	100	100	61	56	78	94	33	38
Total number of inferences (1.1 – 1.5)	28	23	23	21	33	37	8	11
Number of inferences regarding								
1.1 Knowledge	1	0	2	0	5	5	3	2
1.2 Personality traits	7	7	7	9	10	14	0	1
1.3 Disposition	1	4	5	3	5	13	1	4
1.4 Intentions	12	9	4	6	5	3	2	2
1.5 Social category	7	3	5	3	8	2	2	2
Percentage of raters using phase of training (frame of reference)	19	17	28	19	28	31	33	44

Note Presented are the percentages of raters making inferences about the trainee, and the number of inferences made by all raters, in all and per dimension.

Task-specific schemas

Analysis of verbal protocols resulted in 3 major categories reflecting use of task-specific performance schemas (Table 7.1): identification of case-specific cues; identification of particular behaviours as (in)effective, explicitly linked to case-specific cues; and effects of trainee behaviour on the particular patient. These categories represent raters’ comments which do not just focus on discrete aspects of raters’ performance theory, but explicitly and specifically link (in)effectiveness of behaviours and performance to case-specific cues. These features of task-specific performance schemas reflect raters’ efforts to understand the requirements of task-specific performance and the use of ‘task-specific performance theory’ to interpret and evaluate what is happening during the patient encounter.

> *“This patient is very demanding. And then you know that he (the student) cannot get away with this. He (the patient) wants to overrule the doctor’s decision, and then he (the student) will also have to overrule …… and see what he can do for this patient.” (PP21)*

> *“He (the student) says “……..” And this is a very important sentence here. It makes the patient feel welcome, and this is very important for this particular patient because he is feeling rather uncomfortable about not having gone to the doctor earlier.” (PP 27)*

Table 7.3 *Examples of inferences about the trainee, per video case*

1.1 Inferences regarding knowledge:

Cardiocase: definitely adequate knowledge base; knowledge inadequate; he finds it difficult to apply knowledge in clinical practice

Dermacase: I think that he will perform well on knowledge tests

1.2 Inferences regarding personality traits:

Cardiocase: <this trainee is> warm-hearted; sympathetic; timid; friendly; well-behaved; nice person

Dermacase: <this trainee is> rigid; cold-hearted; not empathic; interested

1.3 Inferences regarding disposition:

Cardiocase: <this trainee> adopts a clinical approach towards his patients; adopts an open approach; finds it difficult to discuss patients' feelings and emotions; is too much involved with his own thoughts, as are most young residents; finds it difficult to just sit back and listen to the patient, but he will learn in time;

Dermacase: <this trainee> adopts a clinical approach; listens attentively and reacts to others;

1.4 Inferences regarding intention:

Cardiocase: <this trainee> clearly does not want to make any mistakes with this patient; focuses on adequately handling the biomedical aspects of this patient's problem;

Dermacase: <this trainee> definitely wants to stay in charge; focuses on adequately handling the biomedical aspects of this patient's problem; this trainee is eager to demonstrate that he can handle this

1.5 Inferences regarding social category:

Cardiocase: he clearly just finished his clinical clerkships; he cannot think or act like a general practitioner; he has got the capacity to become a good physician; inexperienced;

Dermacase: he is an authoritarian doctor; he is a technical doctor; robot-like; doesn't seem to take any pleasure in being a doctor; quick, efficient worker

2. Phase of training – frame of reference for judging performance

Well, he is a final year student, so I will have to take this into account, won't I?

Person schema

Table 7.2 presents the percentage of raters that made inferences about the trainee, as well as the kind and number of verbal utterances reflecting inferences, per group and per VC, at T1 and T2. Results show that the majority of raters made inferences about trainees during observation and evaluation of performance, especially in case of salient behaviours (VC1). Table 7.2 also shows that raters were most likely to be involved in making inferences at T1, when they were forming their first impressions. All five types of inference processing described by Klimoski and Donahue (2001) appeared to be present in assessment of trainee performance during single patient encounters. Examples of inferences by different raters and for each of the VCs are presented in Table 7.3.

Rater idiosyncrasy

The results for rater idiosyncrasy are presented in Table 7.4. For each performance dimension the percentage of raters using the dimension is presented, per VC at T1 and T2 respectively. The table shows that (nearly) all raters used the performance dimension "doctor-patient relationship" or at least one of its sub-dimensions in both video cases. For all other (sub-) dimensions, the percentages of raters using the dimension varied (often far from 0% or 100%), indicating considerable between-rater differences in the use of performance theory – i.e. rater

Table 7.4 *Performance theory: use of performance dimensions and performance-related behaviours for experienced raters (Exp) and non-experienced raters (N-exp)*

Performance dimensions	Derma case (VC1)				Cardiocase (VC2)			
	T1		T2		T1		T2	
	Exp (N = 16)	N-exp (N = 12)	Exp (N = 18)	N-exp (N = 16)	Exp (N = 18)	N-exp (N = 16)	Exp (N = 18)	N-exp (N = 16)
Think/act like GP	50 (0/8)	42 (0/5)	56 (0/10)	44 (0/7)	5 (0/1)	6 (0/1)	11 (0/2)	13 (1/1)
Doctor-patient relationship total*	100 (4/60)	92 (1/47)	100 (5/77)	94 (7/81)	94 (71/44)	94 (39/25)	89 (12/51)	82 (18/62)
Establishing / developing rapport	63 (4/14)	75 (1/11)	44 (0/9)	38 (0/9)	72 (28/5)	56 (13/3)	17 (3/1)	13 (1/1)
Demonstrating confidence	13 (0/2)	8 (0/1)	0 (0/0)	6 (0/1)	22 (1/4)	6 (0/2)	22 (0/4)	25 (1/7)
Demonstrating empathic behaviour	31 (0/7)	50 (0/8)	50 (0/10)	44 (0/9)	56 (15/4)	56 (11/2)	0 (0/0)	19 (2/2)
Open approach	44 (0/11)	75 (0/13)	33 (1/8)	44 (0/9)	38 (17/1)	69 (11/4)	0 (0/0)	24 (3/2)
Shared mind 1	50 (0/12)	42 (0/8)	56 (0/14)	69 (0/18)	56 (2/17)	44 (1/10)	44 (1/11)	50 (2/16)
Shared mind 2	0 (0/0)	0 (0/0)	39 (2/13)	63 (6/13)	5 (0/1)	6 (0/1)	39 (3/8)	38 (0/15)
Shared mind 3	0 (0/0)	0 (0/0)	44 (2/13)	63 (1/15)	11 (0/3)	0 (0/0)	61 (5/22)	50 (8/13)
(Bio)medical aspects total**	31 (8/0)	33 (5/1)	67 (12/28)	75 (25/16)	61 (21/2)	19 (3/0)	94 (17/37)	81 (12/32)
History taking	19 (5/0)	17 (2/1)	28 (5/2)	56 (12/1)	50 (16/1)	13 (2/0)	22 (3/5)	19 (7/2)
Physical examination	0 (0/0)	0 (0/0)	28 (0/9)	44 (3/9)	0 (0/0)	0 (0/0)	33 (5/4)	32 (5/10)
Diagnosis / DD	0 (0/0)	0 (0/0)	22 (3/1)	25 (3/2)	11 (1/1)	0 (0/0)	33 (2/5)	38 (1/7)
Patient Management	0 (0/0)	0 (0/0)	61 (1/19)	31 (4/4)	0 (0/0)	0 (0/0)	72 (5/23)	56 (8/13)
Structuring and time management	13 (2/1)	0 (0/0)	17 (1/2)	50 (2/9)	44 (7/5)	19 (0/3)	22 (3/4)	44 (6/7)

Note Presented are percentages of raters using a performance dimension and absolute numbers of verbal utterances (effective / ineffective behaviours) in parentheses, for each group of raters and per VC, at T1 and T2 respectively.

* Doctor-patient relationship total = sum of all verbal utterances within the performance dimension "doctor-patient relationship"

** (Bio)medical aspects total = sum of all verbal utterances within performance dimension "(Bio)medical aspects"

Idiosyncrasy - during assessment of trainee performance. No consistent relationship was found between inter-rater differences and rater expertise.

The numbers of effective and ineffective behaviours per dimension per VC (Table 7.4) reflect between-trainee differences in performance effectiveness. The more balanced pattern of effective and ineffective performance-related behaviours in VC2 reflects the more complex and differentiated behaviours of this particular trainee. In general raters seemed to pay more attention to ineffective behaviours (negative information) compared to effective behaviours at T2, when providing an overall judgment of trainee performance after viewing the entire VC.

Rater expertise and use of task-specific schemas

Results with respect to use of task-specific schemas are presented in Table 7.5. Experienced raters paid significantly more attention to task-specific factors when assessing trainee performance. For the complex cardiocase (VC2), significant between-group differences with respect to the number of task-specific elements of performance (A1+A2+A3) used per rater were found at T1 and T2 (*U* = 77.5, *p* = .02, *ES* = .41 resp.*U* = 86, *p* = .04, *ES* = .35). For the dermacase (VC1), similar and near-significant differences were found at T1 (*U* = 57, p = .07). At T2, significant between-group differences were found for task-specific elements (A1+A2) (*U* = 73, *p* = .01, *ES* = .44). Although statements about task-specific factors, in general, account for a relatively small percentage of all verbal utterances, Table 7.5 clearly shows that statements related to task-specific performance schemas represent a substantial part of the verbal protocols of the more experienced raters.

Discussion

Using theoretical frameworks from social perception research, we sought to better understand underpinnings of work-based assessment outcomes through exploring the content and use of schemas by raters during assessment of trainee performance in single patient encounters. Findings from our study offer indications that raters use different schemas, interactively, when evaluating trainee performance in patient encounters: performance theories, task-specific performance schemas as well as person schemas are used to arrive at judgments about performance. Our results, however, indicate substantial between-rater differences with respect to the use of performance theories (i.e. rater idiosyncrasy) as well as ‘expert-novice’ differences in the use of task-specific performance schemas during performance assessment.

Through the use of think-aloud procedures in actual rating tasks, we were able to establish the dimensions of performance that physician-raters in general practice use to assess performance. The performance dimensions as presented in Figure 7.1 emerged from the analysis of performance ratings by a large number of GP-supervisors (N = 34), who rated performance of two different trainees in two different patient encounters. Dimensions and sub dimensions *together* could be considered to reflect a normative performance theory, or ‘performance schema’, of physician performance in general practice, built upon what ‘raters actually pay attention to and comment upon in practice’.

Table 7.5 *Task-specific schema: use of task-specific performance schemas for experienced (Exp) and non-experienced (N-exp) raters*

	Derma case (VC1)				Cardiocase (VC2)			
	T1		T2		T1		T2	
	Exp (N = 16)	N-exp (N = 12)	Exp (N = 18)	N-exp (N = 16)	Exp (N = 18)	N-exp (N = 16)	Exp (N = 18)	N-exp (N = 16)
A1 Case-specific cues (clinical presentation; patient behaviour; setting consultation)	69 (9.8 / 13.1)	42 (.0 / 8.9)	72 (4.9 / 9.4)	31 (.0 / 5.5)	67 (7.2 / 9.3)	38 (.0 / 9.1)	72 (8.5 / 14.6)	56 (1.1 / 8.2)
A2 Specific trainee behaviours	44 (.0 / 7.1)	17 (.0 / .0)	44 (.0 / 7.1)	13 (.0 / .0)	44 (.0 / 7.3)	13 (.0 / .0)	56 (6.7 / 11.3)	38 (.0 / 5.0)
A3 Effects of trainee behaviours	19 (.0 / .0)	8 (.0 / .0)	22 (.0 / .8)	31 (.0 / 6.1)	44 (.0 / 8.4)	0 (-)	50 (1.6 / 7.9)	31 (.0 / 5.0)
A1-3 Task-specific features total	81 (11.8 /14.2)	50 (3.6 / 9.1)	72 (10.6 / 17.9)	56 (6.3 / 11.8)	78 (12.9 / 12.8)	38 (.0 / 13.8)	78 (20.7 / 23.6)	69 (6.7 / 17.2)

Note Presented are the percentages of raters using task- or event-specific elements of performance, and percentages of statements per protocol in parentheses (median / interquartile range), for each group of raters and per VC, at T1 and T2 respectively.

The results from our study seem to be inconsistent with previous research on work-based assessment indicating that raters have a 1- or 2-dimensional conception of professional competence ('cognitive / clinical' and 'humanistic / (psycho)social') and therefore are unable to discriminate between different competencies or dimensions (Cook et al., 2010; Pulito et al., 2007; Archer et al., 2010). This so-called halo effect is generally attributed to rater error, resulting from global impression formation, categorization or 'stereotyping'. Results from our study clearly show that raters distinguish a fairly large number of different performance dimensions. Our results also show, however, that raters use dimensions interactively when assessing performance. For example, when assessing performance during history taking, physical examination, or patient management, raters assessed not only trainees' ability to adequately handle (bio)medical or 'medico-technical' aspects of the problem, but also their communication and interpersonal as well as time management skills. In other words, the performance theory (or competency framework) used by raters does not map neatly onto the frameworks as displayed on most standardized rating scales which present performance dimensions as strictly separate, distinct entities (e.g. the typical mini-CEX format). True correlations between different performance dimensions may be high, and observed halo effects may – at least partially- be considered as 'true halo' rather than the result of rater incompetence or automatic top-down categorization of trainee performance.

Our findings also show that GP-supervisors differ with respect to the dimensions used in performance evaluation, indicating varying levels of rater idiosyncrasy. Furthermore, raters used different dimensions, depending on what they actually saw during the patient encounter: apparently, not all dimensions are equally relevant or important in all cases. In general, standardized rating scales are designed to represent a given set of performance dimensions (or competencies) in a predefined order, suggesting equal importance of each performance domain. Requiring raters to fill in a rating score for all performance dimensions may therefore hinder accurate depiction of trainee performance. Our findings are in line with findings from Ginsburg et al. (2010) who found that, in evaluations of resident performance, dimensions took on variable degrees of importance, depending on the resident.

The present study confirms findings of expertise research indicating that when handling complex tasks, 'experts' pay more attention to contextual or situation-specific factors before deciding on a plan of action or solution (e.g. Ross et al., 2006). When assessing trainee performance in patient encounters, experienced GP-raters pay (significantly) more attention to task-specific cues. Furthermore, experienced raters seem to be more likely than non-experienced raters to explicitly link task- or case-specific cues to specific trainee behaviours and effects of trainee behaviour on both the patient and the outcome of the patient consultation. Similar results were found in a study on teacher supervision, where experienced supervisors, to a greater extent than non-experienced supervisors, automatically looked for coherence and meaning in teacher behaviours. Experienced supervisors searched for student involvement and effects of teacher behaviours on student learning, rather than focusing exclusively on discrete aspects of teacher behaviours (Kerrins and Cushing, 2000). Our findings thus suggest that experienced raters have more differentiated performance schemas which are activated by task-specific cues. In this respect, our findings are consistent with previous research in industrial and organizational psychology showing that experienced raters are more sensitive to relevant ratee behaviours and have more and more sophisticated performance schemas (Cardy et al., 1987).

Findings from our study clearly indicate that raters start developing person schemas from the first moment they start to observe trainee performance. Raters not only make inferences about knowledge and disposition based on what they know about the trainee (phase of training, for instance), but at least some raters also seem to categorize trainees according to personality judgments and behavioural interpretations. Although our findings show consensus among raters with respect to some inferences about individual trainees, there is also considerable disagreement. These findings are in line with person perception research, which consistently shows that perceivers' <idiosyncratic> interpretive processes may produce sharp differences in person perception (Mohr and Kenny, 2006). In general, people make social inferences spontaneously (Uleman et al., 2008; Macrae and Bodenhausen, 2001) and raters' person schemas -once developed - may guide (selective) attention in subsequent assessments and colour the interpretation of future information. Differences in the way raters form person schemas in work-based assessment contexts may therefore be one of the major factors underlying differences in rating outcomes.

Limitations of our study

This study has several limitations. Since all participants in our study were volunteers they may have been more motivated to carefully assess trainee performance. Together with the experimental setting of our study, this may limit generalization of our findings to raters in 'real life' general practice. Real life settings are most often characterized by time constraints, conflicting tasks and varying rater commitment, which may all impact on rater information processing. Another limitation of our study may be that raters were all selected within one geographical region. As a consequence the normative performance theory that evolved from our data may reflect the structure of assessment tools that are used in training and local GP-supervision practices, and generalization of our results to other regions or disciplines may be limited. Nevertheless, one of the main findings of our study remains that raters show considerable levels of idiosyncrasy, despite rater training and prolonged task experience.

A further limitation is the way we selected experts. We used only years of experience as a measure of expertise; other variables such as actual supervisor performance, commitment to teaching and assessment, or reflectiveness were not measured or controlled for. Time and experience are clearly important variables in acquiring expertise, though. In this respect, our relative approach to expertise is very similar to approaches in expertise research in the domain of clinical reasoning in medicine (Norman, 2006).

In the setting of our experiment, participants were asked to think aloud while providing judgments about others. The task of verbalizing thoughts while filling out a rating scale and providing a performance score may have introduced an aspect of accountability into the rating task, with both experienced and non-experienced raters feeling compelled to retrospectively explain and justify their actions. However, since providing motivations for any performance rating while giving feedback and discussing ratings with trainees is a built-in characteristic of performance evaluation in general practice, our experimental setting comes close to real life task performance, and verbal protocols in our study most likely reflect 'natural' cognitive processing by raters in 'context-free' assessment of performance.

Implications of our study

Results from our study have several implications for WBA practice as well as future research. Firstly, our findings may have implications for rater training. Our findings provide further support for the implementation of 'frame-of reference' (FOR) training as proposed by Holmboe (2008). As indicated before, results of rater training are often disappointing and one of the major reasons may be that, in general, rater training tends to focus on how to use predefined and standardized assessment instruments, ignoring raters' a priori performance theories. As a consequence, transfer of training may be limited. Frame-of-reference training on the other hand asks raters to reflect on their personal methods for evaluating performance, and aims to reduce idiosyncratic rating tendencies through discussing and defining performance dimensions, performance-related behaviours as well as performance levels. FOR training, in other words, establishes a 'shared mental model' or 'shared performance theory' for observing and evaluating performance. In the performance appraisal domain, FOR training has emerged as the most promising rater training approach and it has been successfully applied in field settings (Sulsky and Kline, 2007; Holmboe et al., 2004).

Secondly, our findings may have implications for the way we select raters in the context of WBA. Based on the findings from our study, the use of task-specific performance schemas by more experienced raters may affect feedback given to learners/trainees. The incorporation of contextual cues by experienced raters can result in qualitatively different, more holistic feedback, focusing on a variety of issues and giving meaning to what is happening in the patient encounter by integrating different aspects of performance. Furthermore, research in industrial & organizational psychology indicates that more experienced raters who use more differentiated performance schemas, provide more accurate ratings (e.g. Cardy et al., 1987; Ostroff and Ilgen, 1992). Although we did not aim to investigate the relationship between use of schemas and rating accuracy, our findings definitely put forward the need for further research into effects of rater expertise on accuracy of work-based performance assessment.

Results may also have implications for design of rating scales or rating formats in WBA. As indicated before, correct interpretation of rating scores and usefulness of performance ratings may be compromised by the use of rating scales which do not adequately mirror raters' performance theories. Eliciting "performance theory-in-use", as in our experimental setting or as part of FOR-training procedures, may contribute to development of assessment frameworks and instruments which reflect what experienced practitioners find important in their judgment of trainees. It is to be expected that the use of rating instruments which are in line with raters' natural cognitive processing and competency frameworks will generate more valid and authentic performance ratings, improving usefulness of WBA results.

More importantly, however, we feel that our findings illustrate the importance of narrative, descriptive feedback in WBA. From our findings, it is clear that a simple score on a rating scale is determined by complex and idiosyncratic information processing. Meaningful interpretation of performance scores therefore requires additional narrative comments providing insight into the rater's personal motivations and argumentations. Narrative feedback and comments will thus support credible and defensible decision making about competence achievement. Moreover, narrative feedback -if provided in a constructive way- is the only way to help trainees to accurately identify strengths and weaknesses in performance and to effectively guide trainees' competence development.

Finally, the development and use of person schemas may pose a threat to validity of WBA results (e.g. risk of stereotyping). It is important to realize, however, that schema-based processing in performance assessments is likely to be inevitable: use of schemas helps raters to efficiently process and organize information about ratees. Therefore, efforts to improve WBA should be directed at design of assessment environments in which any unintended effects of schema-based processing are countered. First of all, it seems important that raters are aware of and recognize the processes by which they form impressions of trainee performance. This requires training, feedback and reflection on performance ratings as well as interactions with others involved in the assessment process. More importantly, however, there is recent evidence that social-cognitive processes that underlie judgments (for example the application of stereotypes) are extremely malleable and adaptive to the perceiver's social goals, motives, emotional state and relationships with others (Smith and Semin, 2007). In other words: activation and application of mental representations or knowledge structures such as person schemas, formerly thought to be subconscious and automatic, is influenced by the social context in which judgments have to be made. Based on research in work settings in other domains, effective interventions include allocation of adequate resources (time and money) and providing raters with adequate opportunities to observe and evaluate trainees; ensuring prolonged engagement; holding raters accountable for their decisions; and underscoring mutual interdependence between supervisor and trainee (Operario and Fiske, 2001).

Conclusive remarks

We feel that the findings of our study contribute to a better understanding of the processes underlying work-based assessments in the clinical domain. When assessing performance, raters make use of personal constructs and theories about performance that develop through prolonged task experience. Idiosyncratic use of performance theories as well as person models that raters arrive at during observation and assessment, determine rating outcomes. We conclude that our findings support approaches to work-based assessment which take a social-psychological perspective, considering raters to be active information processors embedded in the social context in which assessment takes place.

Further research should examine whether our findings can be reproduced in other settings and other medical specialties. Important areas for research are the use and development of person schemas and their impact on supervisor behaviour towards trainee, feedback processes and subsequent performance evaluations. Further research questions may address development of performance schemas over time and consequences for assessment instruments, rater training and selection. Clearly, we first and foremost need field studies to investigate how contextual factors influence development and use of schemas by raters, and how they affect rating outcomes.

Ethical approval

Dutch law and the Maastricht University IRB have considered this type of research exempt from ethical review. Pending the installation of a national Medical Education Research Review Board we have taken precautions to protect the interests of all participants (students, patients and GP-supervisors). Participation was voluntary and full confidentiality was guaranteed. Informed consent to record patient encounters and to use recordings for research purposes was obtained from students and patients in the DVDs. All participants were informed about research procedures in writing, and permission to audiotape sessions was obtained. Data were analysed anonymously.

References

Albanese, M.A. (2000). Challenges in using rater judgments in medical education. *Journal of Evaluation in Medical Practice, 6*(3), 305-319.

Archer, J., McGraw, M. & Davies, H. (2010). Assuring validity of multisource feedback in a national programme. *Arch Dis Child, 95*, 330-335.

Arts, J.A.R.M., Gijselaers, W.H. & Boshuizen, H.P.A. (2006). Understanding managerial problem-solving, knowledge use and information processing: Investigating stages from school to the workplace. *Contemporary Educational Psychology, 31*(4), 387-410.

Barnes-Farrell, J.L. (2001). Performance appraisal: Person perception processes and challenges. In: M. London (Ed.), *How People Evaluate Others in Organizations* (pp 135-153). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Borman, W.C. (1987). Personal constructs, performance schemata and "folk theories" of subordinate effectiveness: Explorations in an army officer sample. *Organizational Behavior and Human Decision Processes, 40*, 307-322.

Cardy, R.L., Bernardin, H.J., Abbott, J.G., Senderak, M.P. & Taylor K. (1987). The effects of individual performance schemata and dimension familiarization on rating accuracy. *Journal of Occupational Psychology*, *60*, 197–205.

Chi, M. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences, 6*(3), 271-315.

Cook, D.A., Beckman, T.J., Mandrekar, J.N. & Pankratz, V.S. (2010). Internal structure of mini-CEX scores for internal medicine residents: Factor analysis and generalizability. *Advances in Health Sciences Education*. Doi: 10.1007/s10459-010-9224-9.

Davies, H. (2005). Work based assessment. *BMJ Career Focus, 331*, 88-89.

DeNisi, A.S. (1996). *Cognitive Approach to Performance Appraisal: A Program of Research*. New York: Routledge.

Field, A. (2009). *Discovering statistics using SPSS*. London, etc.: Sage Publications Ltd.

Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K. & Regehr, G. (2010). Toward authentic clinical evaluation: Pitfalls in the pursuit of competency. *Academic Medicine, 85*, 780-86.

Govaerts, M., Van der Vleuten, C., Schuwirth, L. & Muijtjens, A. (2007). Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. *Advances in Health Sciences Education, 12*, 239-260.

Govaerts, M.J.B., Schuwirth, L.W.T., Van der Vleuten. C.P.M. & Muijtjens, A.M.M (2010). Workplace-based assessment: Effects of rater expertise. *Advances in Health Sciences Education*. DOI: 10.1007/s10459-010-9250-7 (e-pub ahead of print).

Green, M. & Holmboe, E. (2010). The ACGME toolbox: Half empty or half full? *Academic Medicine, 85*(5), 787-790.

Hoffman, B., Lance, C.E., Bynum, B & Gentry, W. (2010). Rater source effects are alive and well after all. *Personnel Psychology, 63*, 119-151.

Holmboe, E.S. (2008). Direct observation by faculty. In: E.S. Holmboe & R.H. Hawkins (Eds.), *Practical Guide to the Evaluation of Clinical Competence* (pp. 119-29). Philadelphia: Mosby-Elsevier.

Holmboe, E.S., Hawkins, R.E. & Huot, S.J. (2004). Effects of training in direct observation of medical residents' clinical competence. *Ann Intern Med, 140*, 874-881.

Holmboe, E.S., Sherbino, J., Long, D.M., Swing, S.R. & Frank, J.R. (2010). The role of assessment in competency-based medical education. *Medical Teacher, 32*, 676-682.

Kerrins, J.A. & Cushing, K.S. (2000). Taking a second look: Expert and novice differences when observing the same classroom teaching segment a second time. *Journal of Personnel Evaluation in Education*, *14*(1), 5-24.

Klimoski, R.J. & Donahue, L.M. (2001). Person perception in organizations: An overview of the field. In: M. London (Ed.), *How People Evaluate Others in Organizations* (pp 5-43). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kogan, J.R., Holmboe, E.S. & Hauer, K.E. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. *JAMA, 302*(12), 1316-1326.

Landy, F.J. & Farr, J.L. (1980). Performance rating. *Psychological Bulletin, 87*, 72-107.

Levy, P.E. & Williams, J.R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management*, *30*, 881-905.

Lurie, S.J., Mooney, C.J. & Lyness, J.M. (2009). Measurement of the general competencies of the Accreditation Council for Graduate Medical Education: A systematic review. *Academic Medicine, 84,* 301-309.

Macrae, C.N. & Bodenhausen, G.V. (2001). Social cognition: Categorical person perception. *British Journal of Psychology, 92*, 239-255.

McGaghie, W.C., Butter, J. & Kaye, M. (2009). Observational assessment. In: S.M. Downing & R. Yudkowsky (Eds.), *Assessment in Health Professions Education* (pp. 185-216). New York, NY: Routledge.

Mohr, C.D. & Kenny, D.A. (2006). The how and why of disagreement among perceivers: An exploration of person models. *Journal of Experimental Social Psychology, 42*, 337-349.

Murphy, K.R. & Cleveland, J.N. (1995). *Understanding Performance Appraisal: Social, Organizational and Goal-based Perspectives*. Thousand Oaks, CA: Sage Publications.

Murphy, K.R., Cleveland, J.N., Skattebo, A.L. & Kinney, T.B. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology, 89*(1), 158-164.

Norcini, J.J. (2005). Current perspectives in assessment: The assessment of performance at work. *Medical Education, 39*, 880-89.

Norcini, J. & Burch, V. (2007). Workplace-based assessment as an educational tool. AMEE Guide No. 31. *Medical Teacher, 29*(9), 855-71.

Norman, G., Eva, K., Brooks, L. & Hamstra, S. (2006). Expertise in medicine and surgery. In: K.A. Ericsson, N. Charness, P.J. Feltovich & R.R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 339-354). Cambridge: Cambridge University Press.

Operario, D. & Fiske, S.T. (2001). Causes and consequences of stereotypes in organizations. In: M. London (Ed.), *How People Evaluate Others in Organizations* (pp 45-62). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Ostroff, C. & Ilgen, D.R. (1992). Cognitive categories of raters and rating accuracy. *Journal of Business and Psychology, 7*(1), 3-26.

Pennington, D. C. (2000). *Social cognition*. *Routledge Modular Psychology Series*. London: Routledge.

Pulito, A.R., Donnelly, M.B. & Plymale, M. (2007). Factors in faculty evaluation of medical students' performance. *Medical Education, 41*, 667-675.

Ross, K.G., Shafer, J.L. & Klein, G. (2006). Professional Judgments and "Naturalistic Decision Making". In: K.A. Ericsson, N. Charness, P.J. Feltovich & R.R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 403-420). Cambridge: Cambridge University Press.

Scullen, S.E., Mount, M.K. & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*(6), 956-970.

Smith, E.R. & Collins, E.C. (2009). Contextualizing person perception: Distributed social cognition. Psychological Review, 116(2), 343-364.

Smith, E.R. & Semin, G.R. (2007). Situated social cognition. *Current Directions in Psychological Science, 16*(3), 132-135.

Sulsky, L.M. & Kline, T.J.B. (2007). Understanding frame-of-reference training success: A social learning perspective. *International Journal of Training and Development, 11*(2), 121-131.

Uggerslev, K.L. & Sulsky, L.M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology, 93*(3), 711-19.

Uleman, J.S., Saribay, S.A. & Gonzalez, C.M. (2008). Spontaneous inferences, implicit impressions and implicit theories. *Annu. Rev. Psychol., 59*, 329-360.

Veldhuijzen, W., Ram, P.M., Van der Weijden, T., Niemantsverdriet, S & Van der Vleuten, C.P.M. (2007). Characteristics of communication guidelines that facilitate or impede guideline use: A focus group study. *BMC Family Practice, 8*:31. Doi: 10.1186/1471-2296-8-31

Viswesvaran, O., Ones, D.S. & Schmidt, F.L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*(5), 557-574.

Williams, R.G., Klamen, D.A. & McCaghie, W.C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine, 15*(4), 270-292.

CHAPTER 8

Summary and General Discussion

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.

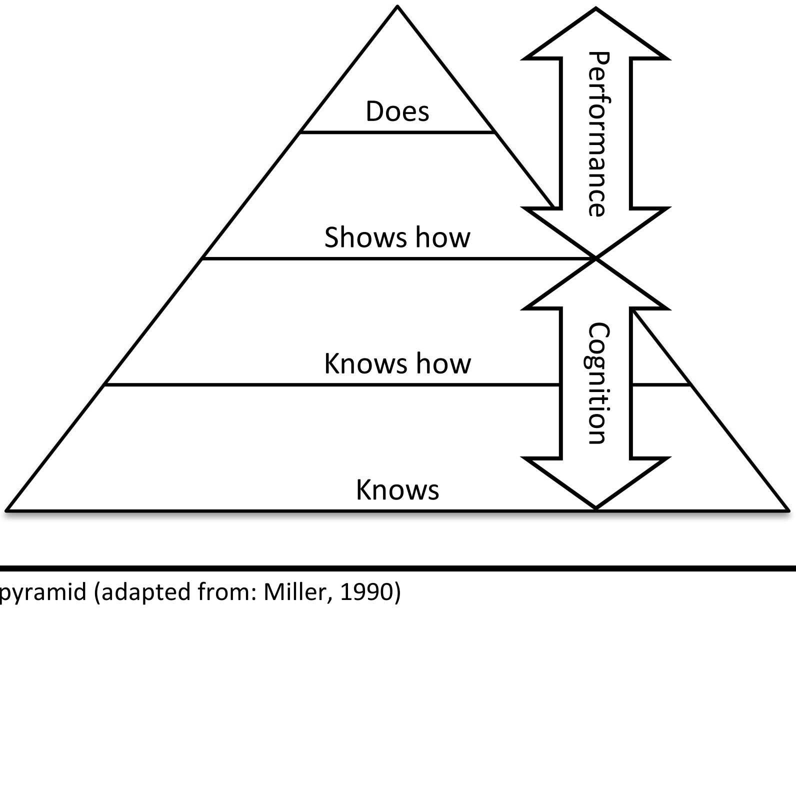


Figure 8.1 Miller' s pyramid (adapted from: Miller, 1990)

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 Rajaratanam, 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 ratee 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 ratee 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

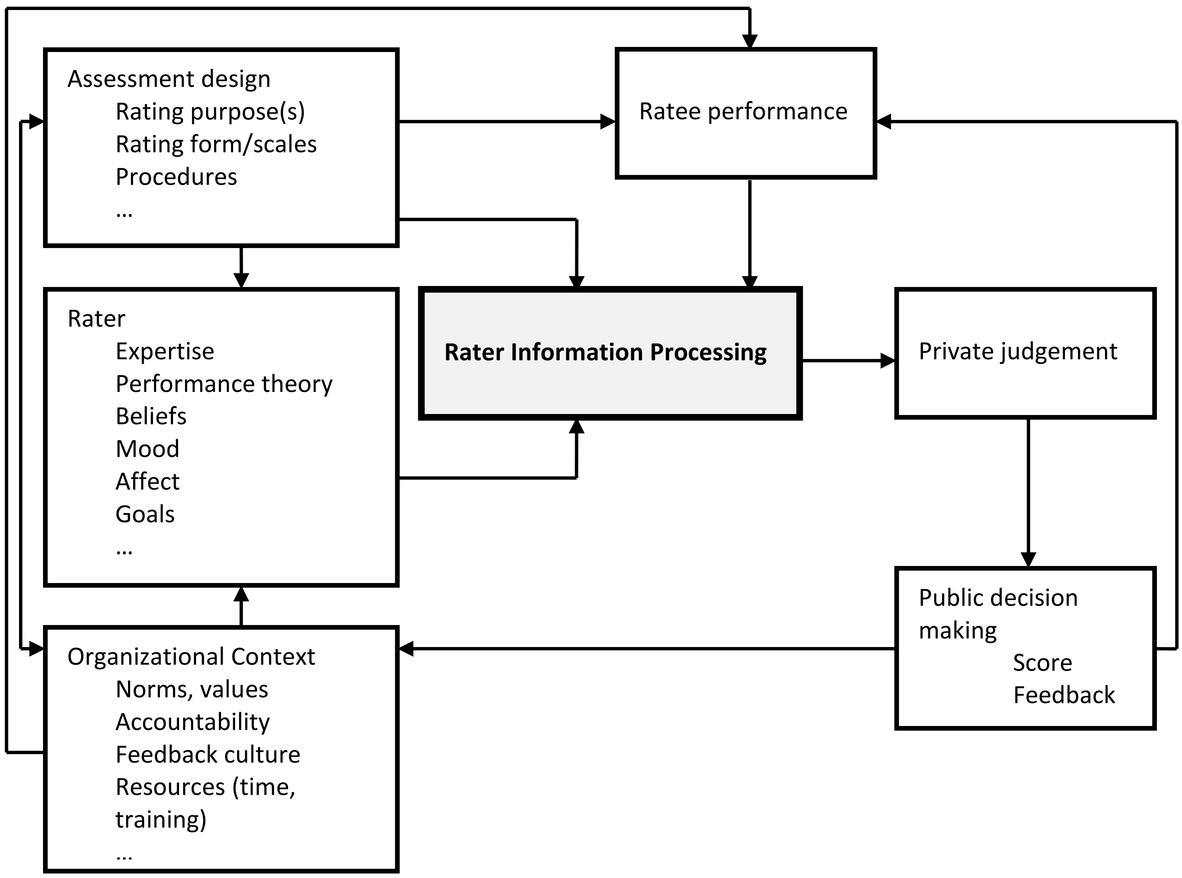


Figure 8.2 Model of the performance assessment process (adapted from DeNisi, 1996)

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

Albanese, M.A. (2000). Challenges in using rater judgments in medical education. *Journal of Evaluation in Medical Practice, 6*(3), 305-319.

Brooks, M.A. (2009). Medical education and the tyranny of competency. *Perspectives in Biology and Medicine, 52(*1), 90-102.

Brutus, S. (2010). Words versus numbers: A theoretical exploration of giving and receiving narrative comments in performance appraisal*. Human Resource Management Review, 20*(2), 144-157.

Canavan, C., Holtman, M.C., Richmond, M. & Katsufrakis, P.J. (2010). The quality of written comments on professional behaviors in a developmental multisource feedback program. *Academic Medicine, 85*, S106–S109.

Cardy, R.L., Bernardin, H.J., Abbott, J.G., Senderak, M.P. & Taylor K. (1987). The effects of individual performance schemata and dimension familiarization on rating accuracy. *Journal of Occupational Psychology*, *60*, 197–205.

Chi, M.T.H. (2006). Two approaches to the study of experts' characteristics. In: K.A. Ericsson, N. Charness, P.J. Feltovich & R.R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 21-30). Cambridge: Cambridge University Press.

Choudhry, N.K., Fletcher, R.H. & Soumerai, S.B. (2005). Systematic review: The relationship between clinical experience and quality of health care. *Ann Intern Med, 142*, 260-273.

Clauser, B. E., Harik, P. & Margolis, M. J. (2006). A multivariate generalizability analysis of data from a performance assessment of physicians' clinical skills. *Journal of Educational Measurement,* 43, 173– 191.

Coles, C. (2002). Developing professional judgment. *The Journal of Continuing Education in the Health Professions*,*22*, 3–10.

Colliver, J.A., Vu. N.V., Markwell. S.J. & Verhulst, S.J. (1991). Reliability and efficiency of clinical competence assessed with five performance-based examinations using standardized patients. *Medical Education, 25*, 303-10.

Cooke, M.C., Irby, D.M. & O'Brien, B.C. (2010). *Educating Physicians. A Call For Reform Of Medical School And Residency.* San Francisco: Jossey-Bass.

Cunnington, J.F.W., Neville, A.J. & Norman, G.R. (1997). The risk of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education 1,* 227–233.

Davis, D.A. (2009). How to help professionals maintain and improve their knowledge and skills: Triangulating best practices in medicine. In: K.A. Ericsson (ed.) *Development of Professional Expertise. Toward Measurement of Expert Performance and Design of Optimal Learning Environments* (pp. 180-202). Cambridge, etc.: Cambridge University Press.

Davis, D.A., Mazmanian, P.E., Fordis, M., Harrison, V.R., Thorpe, K.E. & Perrier, L. (2006). Accuracy of physician self-assessment compared with observed measures of competence: A systematic review. *JAMA, 296*, 1094-1102.

DeNisi A.S. & Peters, L.H. (1996). Organization of information in memory and the performance appraisal process: Evidence from the field. *Journal of Applied Psychology, 81*(6), 717-737.

DeNisi, A.S. (1996). *Cognitive Approach to Performance Appraisal: A Program of Research*. New York: Routledge.

DeNisi, A.S., Robbins, T. & Cafferty, T.P. (1989). Organization of information used for performance appraisals: Role of diary-keeping. *Journal of Applied Psychology 74*(1), 124-129.

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

Duffy, F.D. & Holmboe, E.S. (2006). Self-assessment in lifelong learning and improving performance in practice : Physician know thyself. *JAMA, 296* (9), 1137-1139.

Epstein, R.M. & Hundert, E.M. (2002). Defining and assessing professional competence. *JAMA*, *287*(2), 226-235.

Erdogan, B., Kraimer, M.L. & Liden, R.C. (2001). Procedural justice as a two-dimensional construct. An examination in the performance appraisal context. *Journal of Applied Behavioural Science 37*(2), 205-222.

Ericsson, K.A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 79*(10), S70-S81.

Ericsson, K.A. (2009). Enhancing the development of professional performance: Implications from the study of deliberate practice. In: K.A. Ericsson (Ed.) *Development of Professional Expertise. Toward Measurement of Expert Performance and Design of Optimal Learning Environments* (pp. 405-431). Cambridge: Cambridge University Press.

Eva, K.W. & Regehr, G. (2005). Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine, 80*, S46-54.

Ferris, G.R., Munyon, T.P., Basik, K. & Buckley, M.R. (2008). The performance evaluation context: Social, emotional, cognitive, political and relationship components. *Human Resource Management Review, 18*, 146-163.

Gordon, M.J. (1991). A review of the validity and accuracy of self-assessments in health professions training. *Academic Medicine, 66*, 762–769.

Hager, P., Gonczi, A. & Athanasou, J. (1994). General issues about assessment of competence. *Assessment & Evaluation in Higher Education*, *19*(1), 3-16.

Hattie, J. & Timperley, H. (2007). The power of feedback*. Review of Educational Research, 77*(1), 81-112.

Hawe, E. (2003). It's pretty difficult to fail: The reluctance of lecturers to award a failing grade. *Assessment and Evaluation in Higher Education 28*(4), 371–382.

Hodges, B. (2006). Medical education and the maintenance of incompetence. *Medical Teacher, 28*(8), 690-696.

Hodges, B., Regehr, G., McNaughton, N., Tiberius, R. & Hanson, M. (1999). OSCE checklists do not capture increasing levels of expertise. *Academic Medicine, 74,* 1129–1134.

Howley, L.D. (2004). Performance assessment in medical education: Where we've been and where we're going. *Eval Health Prof, 27*, 285-303.

Jones, A. (1999). The place of judgement in competency-based assessment. *Journal of vocational education and training*, *(51)*1. 145-160.

Kane, M. (2006). Validation. In: R.L. Brennan (Ed.) *Educational Measurement* (pp.621-694). Westport CT: American Council on Education/Praeger.

Kane, M., Crooks, T. & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *Summer 1999*, 5-17.

Keller, L.A., Clauser, B.E. & Swanson D.B. (2010). Using multivariate generalizability theory to assess the effect of content stratification on the reliability of a performance assessment. *Advances in Health Sciences Education*. DOI 10.1007/s10459-010-9233-8.

Kerrins, J.A. & Cushing, K.S. (2000). Taking a second look: Expert and novice differences when observing the same classroom teaching segment a second time. *Journal of Personnel Evaluation in Education*, *14*(1), 5-24.

Klimoski, R.J. & Donahue, L.M. (2001). Person perception in organizations: An overview of the field. In: M. London (Ed.), *How People Evaluate Others in Organizations* (pp 5-43). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kluger, A.N. & DeNisi, A.S. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254-284.

Korthagen, F. & Vasalos, A. (2005). Levels in reflection: Core reflection as a means to enhance professional growth. *Teachers and Teaching: theory and practice, 11*(1), *47–71.*

Kramer, A., Muijtjens, A., Jansen, K., Düsman, H, Tan, L. & Van der Vleuten, C. (2003). Comparison of a rational and an empirical standard setting procedure for an OSCE. *Medical Education*, *37*, 132–139.

Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessment. *Journal of Personality and Social Psychology, 77*(6), 1121-1134.

Lance, C.E., Hoffman B.J., Gentry, W.A. & Baranik, L.E. (2008). Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Human Resource Management Review, 18*, 223-232.

Levy, P.E. & Williams, J.R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management*, *30*, 881-905.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, *86*(2), 255-264.

Luthans, F. & Peterson, S.J. (2003). 360-degree feedback with systematic coaching: Empirical analysis suggests a winning combination. *Human Resource Management, 42*(3), 243-256.

McDowell, L. (1995). The impact of innovative assessment on student learning. *Innovations in Education and Training International, 32*(4), 302-313.

Mero, N.P. & Motowidlo, S.J. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology 80*(4), 517-524.

Mero, N.P., Motowidlo, S.J. & Anna, A.L. (2003). Effects of accountability on rating behavior and rater accuracy. *Journal of Applied Social Psychology 33*(12), 2493-2514.

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

Mohr, C.D. & Kenny, D.A. (2006). The how and why of disagreement among perceivers: An exploration of person models. *Journal of Experimental Social Psychology, 42*, 337-349.

Murphy, K.R. & Cleveland, J.N. (1995). *Understanding performance appraisal. Social, Organizational and Goal-based Perspectives*. Thousand Oaks, CA: Sage Publications.

Norcini, J.J. & McKinley, D.W. (2007). Assessment methods in medical education. *Teaching and Teacher Education, 23*, 239-250.

Overeem, K., Lombarts, M. Arah, O., Klazinga, N., Grol, R. & Wollersheim, H. (2010). Three methods of multi-source feedback compared: A plea for narrative comments and coworkers' perspectives. *Medical Teacher, 32*(2), 141-147.

Regehr, G., MacRae, H., Reznick, R.K. & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine 73*(9), 993–997.

Reznick, R.K. & Rajaratanam, K. (2000). Performance-based assessment. In L.H. Distlehorst, G.L. Dunnington & J.R. Folse (Eds.), *Teaching and learning in medical and surgical education: Lessons Learned for the 21st Century* (pp. 237–244). Mahwah: Lawrence Erlbaum Associates.

Sargeant, J., Mann, K., Van der Vleuten, C. & Metsemakers, J. (2008). "Directed" self-assessment: Practice and feedback within a social context. *Journal of Continuing Education in the Health Professions, 28*(1), 47-54.

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes, 53*, 252-266.

Shute, V.J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189.

Singer, P.A., Robb, A., Cohen, R., Norman, G. & Turnbull, J. (1996). Performance-based assessment of clinical ethics using an Objective Structured Clinical Examination. *Academic Medicine 71*(5), 495–498.

Smither, J. W. & Walker, A. G. (2004). Are the characteristics of narrative comments related to improvement in multirater feedback over time? *Journal of Applied Psychology, 89*, 575−581.

Struyven, K., Dochy, F. & Janssens, S. (2003). Students' perceptions about new modes of assessment in higher education: A review. In: M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising New Modes of Assessment: In Search of Qualities and Standards* (pp. 171-223). Dordrecht: Kluwer Academic Publishers.

Taylor, M.S., Tracy, K.B., Renard, M.K., Harrison, J.K & Carroll, S.J. (1995). Due process in performance appraisal: A quasi-experiment in procedural justice. *Administrative Science Quarterly, 40*, 495-523.

Ten Cate, O. (2007). Medical education in the Netherlands. *Medical Teacher, 29*, 752-757.

Teunissen, P.W. (2009). *Unravelling learning by doing. A study of workplace learning in postgraduate medical education*. PhD-thesis. Amsterdam: Vrije Universiteit Amsterdam.

Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In: M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising New Modes of Assessment: In Search of Qualities and Standards* (pp 55-87). Dordrecht: Kluwer Academic Publishers.

Tziner, A., Murphy, K.R. & Cleveland, J.N. (2001). Relationships between attitudes towards organizations and performance appraisal systems and rating behavior. *International Journal of Selection and Assessment, 9*(3), 226-239.

Uggerslev, K.L. & Sulsky, L.M. (2008). Using frame-of-reference training to understand the Implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology, 93*(3), 711-19.

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

Vleuten van der, C.P.M., Schuwirth, L.W.T., Scheele, F., Driessen, E.W. & Hodges, B. (2010). The assessment of professional competence: Building blocks for theory development. *Best Practices & Research Clinical Obstetrics and Gynaecology.* Doi: 10.1016/j.bpobgyn.2010.04.001.

Walker, A.G. & Smither, J.W. (1999). A five-year study of upward feedback: What managers do with their results matters. *Personnel Psychology, 52* (2), 393-423.

Watling, C.J., Kenyon, C.F., Zibrowski, E.M., Schulz, V., Goldszmidt, M.A., Sing, Maddocks, H.L. & Lingard, L. (2008). Rules of engagement: Residents' perceptions of the in-training evaluation process. *Academic Medicine, 83*(10 Suppl), S97-100.

Yudkowsky, R. (2009). Performance tests. In: S.M. Downing & R. Yudkowsky (Eds.), *Assessment in Health Professions Education* (pp. 217-244). New York, NY: Routledge.

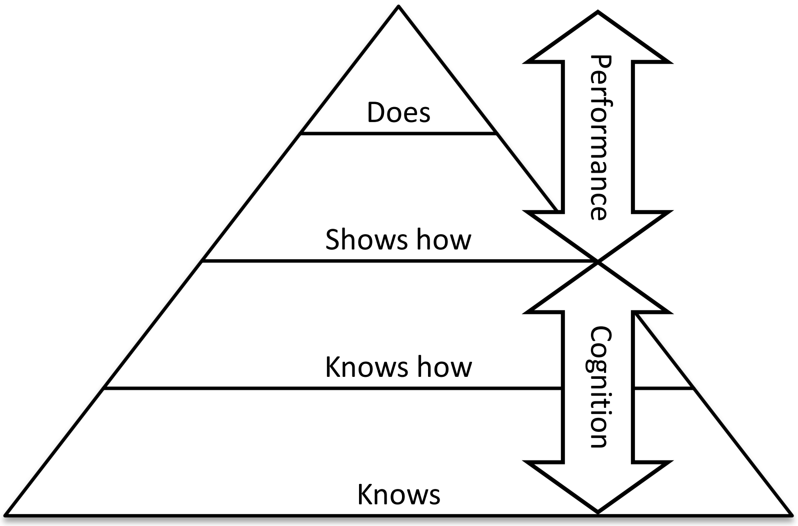
SAMENVATTING

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.



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 toetsscores werd onderzocht met behulp van generaliseerbaarheidstudies. Tevens is nagegaan hoe de toetsbetrouwbaarheid - en daarmee toetsefficiëntie - beïnvloed wordt door a) het gebruik van globale beoordelingsschalen in plaats van traditionele taakspecifieke 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 tot een toename van toetsbetrouwbaarheid (generaliseerbaarheidscoëfficiënt = 0.61), en deze kan 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 frequente (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. Frequente documentatie van feedback en evaluatie van performance aan de hand van helder gedefinieerde criteria en standaarden stuurt de ontwikkeling van professionele competentie in de gewenste richting, en verhoogt gevoelens van self-efficacy gedurende de stage. Systematische en gestructureerde evaluatie van performance, en 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 numerieke 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.

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 werkplek, gecombineerd kunnen en moeten worden. Deze bevindingen ondersteunen recente ontwikkelingen in de medisch-specialistische vervolgopleidingen, 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 werkplek. 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 werkplek. Expertoordelen ('professional judgements') zijn inherent aan beoordelen van performance op de werkplek. Onderzoek, met name vanuit psychometrisch perspectief, toont aan dat betrouwbaarheid en validiteit van werkplekbeoordelingen 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 werkplekbeoordelingen 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 werkplekbeoordelingen. 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. Beargumenteerd wordt 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 informatieverwerving, 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 suggereert 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.

Exploratie van de schema's die gebruikt worden bij het beoordelen van performance in een authentiek consult ondersteunt ons pleidooi voor een sociaal-cognitieve benadering van 'performance assessment', waarin 'performance assessment' wordt beschouwd als een specifieke toepassing van 'social perception' voor specifieke doeleinden, en beoordelaars worden gezien als 'social perceivers'. Bij directe observatie en evaluatie van performance maken beoordelaars gebruik van verschillende kennisstructuren (schema's). Persoonlijke normatieve performance-schema's (ook wel performance theorieën genoemd), taakspecifieke performance-schema's en persoonschema's worden interactief gebruikt bij de totstandkoming van oordelen over performance. Analyse van de verbale protocollen stelde ons in staat een collectieve, normatieve performance theorie te ontwikkelen voor het beoordelen van consultvoering in de huisartspraktijk. In deze theorie worden 17 verschillende, maar met elkaar samenhangende dimensies van performance onderscheiden. Resultaten laten echter ook zien dat er aanzienlijke verschillen bestaan tussen beoordelaars voor wat betreft de performance-dimensies die worden meegenomen in de totstandkoming van het oordeel. Deze bevinding bevestigt het (onvermijdelijk) idiosyncratische karakter van 'performance assessments', als gevolg van verschillen in werkervaring, ervaring en training als supervisor-beoordelaar; en persoonlijke opvattingen over begeleiden c.q. beoordelen. Onze bevindingen met betrekking tot het ontstaan en gebruik van persoonschema's vormen mede een verklaring voor de soms grote verschillen in uitkomsten van werkplekbeoordelingen bij vergelijking van verschillende beoordelaars.

De in hoofdstuk 6 en 7 gepresenteerde bevindingen hebben implicaties voor selectie en inzet van beoordelaars; voor de manier waarop beoordelaars worden getraind, en voor de instrumenten en procedures in 'performance assessment'.

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.

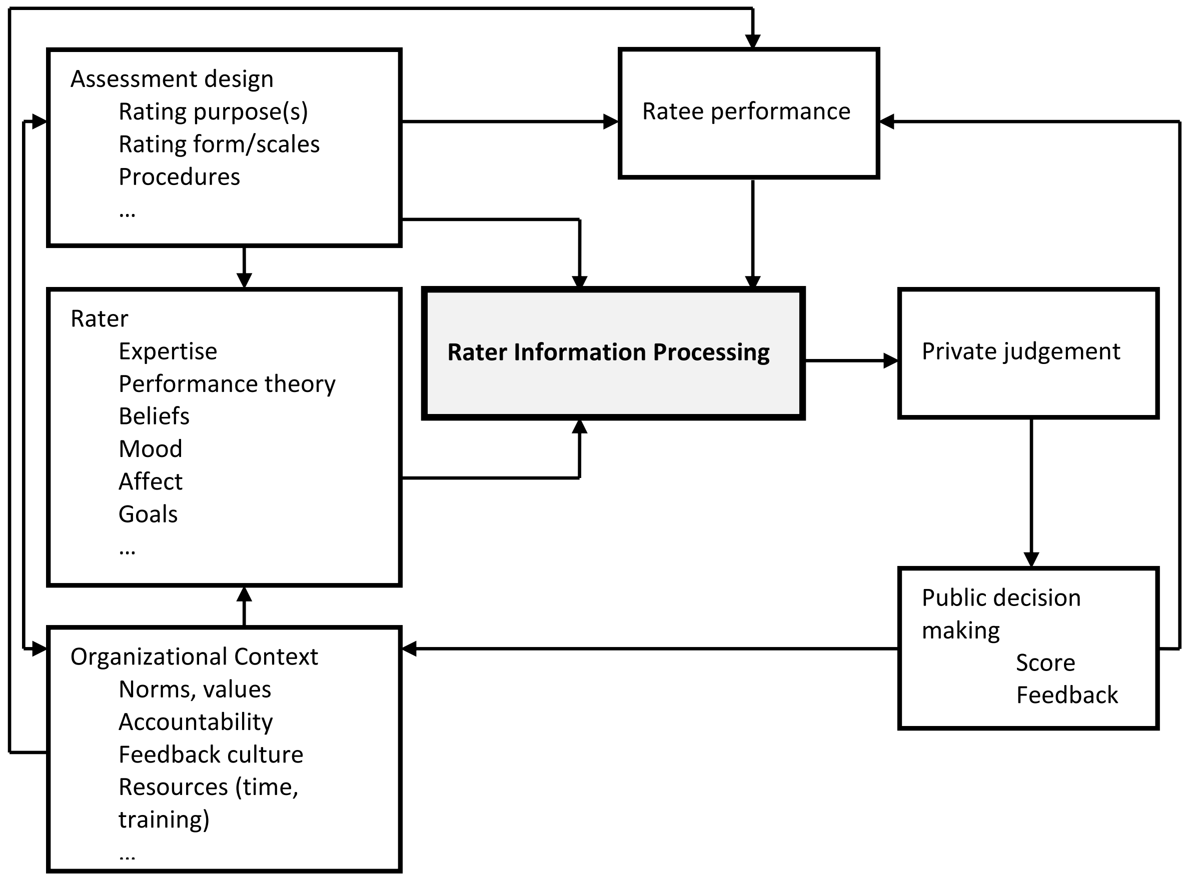
Het eerste thema richt zich op de rol van performance assessments in een *programma* van toetsen in competentiegestuurde opleidingen. Analoog aan recente ontwikkelingen in medische opleidingscurricula wordt gepleit voor een 'Z'-vormig programma van toetsen, dat gekenmerkt wordt door vroege introductie van toetsing van performance. Gestandaardiseerde vormen van performance assessment worden daarbij in toenemende mate vervangen door het toetsen van performance in authentieke situaties, in de authentieke context. Een dergelijke geleidelijke, maar vroeg ingezette overgang naar kwalitatief hoogstaande, longitudinale toetsing van performance op de werkplek kan bijdragen aan de vorming en instandhouding van een cultuur waarin lerenden c.q. professionals in alle fasen van het opleidingscontinuüm worden gestimuleerd tot het vragen om en leren van feedback ten behoeve van performanceverbetering. Belangrijke randvoorwaarden daarbij zijn inbedding van toetsing in leer- en werkprocessen; weloverwogen integratie van formatieve en summatieve toetsfuncties en psychologische veiligheid –zowel voor beoordeelde als beoordelaar. Het creëren van kleinschalige 'communities of learning and practice', met duidelijke taken en 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.

Expertoordelen zijn inherent aan beoordelen van performance (thema 3). Op basis van de studies zoals beschreven in dit proefschrift en onderzoeksbevindingen uit de arbeids- en organisatiepsychologie beargumenteren wij dat expertoordelen in performance assessments vergelijkbaar zijn met ‘professional judgement’ zoals dit gedefinieerd wordt voor andere domeinen. Het beoordelen van performance in de dagelijkse setting van de beroepspraktijk is een complexe taak, uit te voeren onder tijdsdruk (meestal), in een complexe omgeving waarin soms tegenstrijdige belangen een rol spelen. Kennis en ervaring van de beoordelaar, zowel vakinhoudelijk als educatief, zijn van invloed op oordeels- en besluitvorming over performance. Uiteindelijke beslissingen geven vervolgens niet altijd het ‘*meest juiste of accurate*’ oordeel weer, maar wel wat beschouwd wordt als het *‘beste’* oordeel in betreffende situatie, gegeven de omstandigheden. Training, maar vooral relevante ervaring, ‘deliberation’ en ‘deliberate practice’ zijn noodzakelijk voor de ontwikkeling van professional judgement en expertise bij beoordelaars.

Actieve betrokkenheid van beoordeelden c.q. lerenden in het beoordelingsproces, bijvoorbeeld in de vorm van zelfbeoordelingen, is een belangrijke voorwaarde voor acceptatie en gebruik van feedback – en dus voor leren c.q. performanceverbetering. In ‘performance assessments’ zijn oordelen over performance dus niet alleen afkomstig van de supervisoren-beoordelaars, maar ook van de lerenden c.q. beoordeelden. Hoewel onderzoeksresultaten tamelijk consistent aantonen dat mensen niet of nauwelijks in staat zijn een goed oordeel te geven over hun eigen functioneren, zijn er ook aanwijzingen dat het vergelijken van zelfbeoordelingen met beoordelingen en feedback van anderen kan aanzetten tot reflectie en verbetering van performance –mits er sprake is van begeleiding-coaching en een veilige leeromgeving. Het incorporeren van ‘begeleide’ zelfbeoordeling in toetsing van performance, al vroeg in de opleiding, kan op deze manier bevorderen dat studenten zich ontwikkelen tot professionals die in staat zijn tot het formuleren van gefundeerde oordelen (‘informed judgements’) over het eigen functioneren, als basis voor levenslang leren.

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



Figuur 2 Vereenvoudigd model van het performance assessment proces (naar DeNisi, 1996)

beoordeling van performance plaatsvindt. Een groot aantal factoren, waaronder interpersoonlijke, sociale, politieke, emotionele en cognitieve, speelt een rol bij totstandkoming van uitspraken over performance. Figuur 2 presenteert een vereenvoudigd model van onze benadering van (performance) assessment. Factoren die bepalend zijn voor succesvolle implementatie van 'performance assessments' op de werkvloer zijn 'accountability', veiligheid en wederzijds vertrouwen, gebaseerd op eerlijkheid, openheid en rechtvaardigheid ('due process'). Principes van 'accountability' en 'due process' liggen ten grondslag aan een feedbackcultuur waarin het vragen om feedback, reflectie op het eigen functioneren en gebruik van feedback voor performanceverbetering vanzelfsprekend is.

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.

Curriculum vitae

Marjan Govaerts was born in Beek (L.), the Netherlands, on February 11th 1959.

After graduating high school (Gymnasium-B, Maastricht), she studied medicine at the Faculty of Medicine, Maastricht University. She graduated as a medical doctor in 1982. After graduation, she joined the Department of Educational Development and Research at the Faculty of Medicine, Maastricht, and worked on several educational development projects, among which the evaluation of clinical clerkships. She decided to continue her career in the field of education and since 1985 she has been working at various educational institutes in Higher Education, as a teacher and educational consultant. In 1996 she also carried out an evaluative investigation into the 'recognition and admission procedures for foreign doctors', on the authority of the Department of Health, Welfare and Cultural Affairs, the Netherlands.

Before she started her current job she was involved in curriculum development, programme evaluation, student assessment and staff development in a problem-based curriculum for midwifery education, at the Faculty of Midwifery Education and Studies, Maastricht. During that time she started work on her PhD thesis. In 2005, she returned to the Department of Educational Development and Research at the Faculty of Health, Medicine and Life Sciences of Maastricht University. Main activities are centered around (student-)assessment. During the past two years she has been involved in design and implementation of e-portfolios and work-based assessment programmes in postgraduate medical training. Her interests are in competency-based education and assessment, and more specifically in work-based assessment, in-training assessment programmes and expertise development in professional education.

Her work has been published in national and international peer-reviewed journals. In 2010 she received the 'New Investigator Award' from the American Educational Research Association, Division I (Education in the Professions), for her paper on cognition-based approaches towards performance assessment.

Marjan is married to Wim Gijselaers and has two children, Maartje and Sybren.

Dankwoord

De ideeën voor dit proefschrift zijn ontstaan vanuit verbazing en verwondering, vanuit ergernis en frustratie, vanuit zoeken naar antwoorden op praktische problemen en vanuit academische nieuwsgierigheid. Dit proefschrift is, kortom, het resultaat van jarenlange ervaring in het hoger onderwijs en het leerproces dat daar onlosmakelijk mee verbonden is.

Dit proefschrift had ik niet kunnen schrijven zonder de kansen die mij geboden werden, de adviezen die mij gegeven zijn, en de vele hulp en ondersteuning die ik gekregen heb. Het is onmogelijk om iedereen te noemen die - direct of indirect - aan de totstandkoming van dit proefschrift heeft bijgedragen. Begeleiders, collega's en oud-collega's, familie en vrienden, studenten en proefpersonen: ik ben jullie dankbaar!

Mijn begeleiders, Cees, Lambert en Arno: drie uiteenlopende persoonlijkheden, een perfecte mengeling van grote lijn en oog voor detail; van stimulerend enthousiasme en kritische reflectie, en met een ongelooflijk brede vakinhoudelijke expertise. Heel veel dank voor jullie inbreng.

Alle collega's en oud-collega's: jullie leerden mij wat werken in een onderwijsorganisatie betekent. Dank voor jullie collegialiteit, jullie steun en jullie aanmoedigingen om door te zetten, even dat praatje bij de koffie, of dat onverwachte belangstellende mailtje ….

Margje van de Wiel, Annerien Pin, Mieke Clement: jullie hielpen mij over de muren van mijn eigen kennis te kijken. Dat maakt onderzoek doen zó veel leuker en spannender! Esther Bakker, Marlies Schillings, Meike Elferink, Susan van der Vleuten, Hetty Snellen-Balendong, Guus Smeets en Ron Hoogenboom: zonder de door jullie geleverde bijdrage aan dataverzameling en data-analyse was dit proefschrift nog lang niet af geweest. Mereke Gorsira: jouw inbreng maakt van elk stuk iets mooi(er)s. Thanks!

Mijn familie en vrienden: ik maak te weinig tijd voor jullie, maar gelukkig wijzen jullie me met enige regelmaat op de betrekkelijkheid van een en ander, of zoals Konijn het treffend verwoordt: "Er zijn van die dagen dat het niets voorstelt dat je Februari kunt spellen" –dan is het alleen maar belangrijk dat je weet dat de mensen om je heen er voor je zijn. Dank jullie wel voor jullie steun.

Wim, Maartje, Sybren: jullie bijdrage en mijn dank daarvoor kan ik niet in woorden uitdrukken. Jullie zijn het centrum van mijn bestaan.

SHE dissertation series

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

Stalmeijer, R. (2011). Evaluating clinical teaching through cognitive apprenticeship.

Veldhuizen, W. (2011). Challenging the patient centered paradigm: designing feasible guidelines for doctor-patient communication.

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

Van Mook, W. (2011). Teaching and assessment of professional behaviour: Rhetoric and reality.

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

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

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

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

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

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

Niessen, Th. (2007). Emerging epistemologies: Making sense of teaching practice.

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

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

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

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

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

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

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

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