Advances in Health Sciences Education (2007) 12:239–260
DOI 10.1007/s10459-006-9043-1

Broadening Perspectives on Clinical Performance Assessment: Rethinking the Nature of In-training Assessment

MARJAN J. B. GOVAERTS, CEES P. M. VAN DER VLEUTEN, LAMBERT W. T. SCHUWIRTH and ARNO M. M. MUIJTJENS
*Department of Educational Development and Research, Faculty of Medicine, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, The Netherlands (*author for correspondence, E-mail: marjan.govaerts@educ.unimaas.nl; Phone: +31-43-3885746; Fax: +31-43-3885779)*

(Received 28 September 2005; accepted in October 2006)

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 standardisation 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.

Key words: clinical competence, clinical education, clinical ratings, competence assessment, educational measurement, in-training assessment, performance assessment, rating process

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; Prescott et al., 2002; Turnbull and van Barneveld, 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 (Sloan et al., 1995; van der Vleuten et al., 2000). 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 (Verhulst et al., 1986; Ramsey et al., 1993; Silber et al., 2004). 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 (Littlefield et al., 1991; Noel et al.,

1992; Gray, 1996; van Barneveld, 2005). 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. Hull et al., 1995; Kahn et al., 2001; 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 van Barneveld, 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 (Judge and Ferris, 1993; Murphy and Cleveland, 1995).

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 (Fiske and Taylor, 1991; Hodgkinson, 2003; Hogg, 2003).

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 (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 also suggest that non-analytic reasoning is very successful in terms of diagnostic accuracy – even

among relative novices, although its effectiveness increases with medical expertise (Coderre et al., 2003). Research findings also show that excessive 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, 2004).

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 (Cardy et al., 1987; Kozlowski and Mongillo, 1992; Lievens, 2001). 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 (Zedeck, 1986; Fiske and Taylor, 1991; Lance et al., 1994; Lievens, 2001).

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 (Sanchez and DeLaTorre, 1996; Middendorf and Macan, 2002). 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, a 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 (Rothman and Schwarz, 1998; Siemer and Reisenzein, 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; McDowell, 1995; Crooks, 1998; McIlroy et al., 2002). 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; Hawe, 2003; Williams et al., 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 et al., 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 (Cleveland et al., 1989; Murphy and Cleveland, 1995). 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 (Hauenstein, 1992; Mero et al., 2003). 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 et al. (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 (Harris, 1994; Hawe, 2003).

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 cutscore 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 cutscores (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; Taylor et al., 1995; Erdogan et al., 2001; Piggot-Irvine, 2003). 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 (Delandshere and Petroski, 1998; Johnston, 2004) 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, focussing 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?

Acknowledgements

The authors would like to thank Mereke Gorsira for critically reading and correcting the English manuscript.

References

Barneveld, C.van (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*: 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* New Jersey: Hillsdale.

Clauser, B.E. & Schuwirth, L.W.T. (2002). The use of computers in assessment. In: G.R. Norman, C.P.M. Vleutenvan der & 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.

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

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, C.van der, Schuwirth, L., Tartwijk, J.van & 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.

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

Fiske, S.T. & Taylor, S.E. (1991). *Social Cognition* 2 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, C.P.M.van der, 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**: 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, M. III (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: M.A. Hogg *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. *Journal of Applied Psychology* **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.

van Luijk, 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, A.J.J.A. Scherpbier & R.P. Zwierstra (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, J.F. Wergin & 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**: 1762–1771.

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**: 757–765.

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. Vleutenvan der & 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 perfromance (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 & 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 Processses* **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, C.P.M.van der (2005). Quality issues in judging portfolios: implications for organizing teaching portfolio assessment procedures. *Studies in Higher Education* **30**(5): 595–610.

Turnbull, J. & Barneveld, C.van (2002). Assessment of clinical performance: in-training evaluation. In: G.R. Norman, C.P.M. Vleutenvan der & 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**: 132–134.

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

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

Vleuten, C.P.M.van der, 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 envrionmental 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 center 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 Standards: Concepts, Methods and Perspectives*, pp. 19–53. Mahwah NJ: Lawrence Erlbaum Associates.