Adv in Health Sci Educ (2011) 16:151–165
DOI 10.1007/s10459-010-9250-7

Workplace-based assessment: effects of rater expertise

M. J. B. Govaerts · L. W. T. Schuwirth · C. P. M. Van der Vleuten · A. M. M. Muijtjens

Received: 1 June 2010 / Accepted: 16 September 2010 / Published online: 30 September 2010

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 behaviors. 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.

Keywords Clinical education · Cognition-based assessment models · Competence assessment · Performance assessment · Professional education · Professional judgment · Rater expertise · Rating process · Workplace-based assessment

M. J. B. Govaerts (✉) · L. W. T. Schuwirth · C. P. M. Van der Vleuten · A. M. M. Muijtjens
FHML, Department of Educational Research and Development, Maastricht University,
P.O. Box 616, 6200 MD Maastricht, The Netherlands
e-mail: marjan.govaerts@maastrichtuniversity.nl

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 behavior 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 behavior); 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 behavior 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 behavior 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 behavior 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 behaviors, but more time when more complex behaviors 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 behaviors. 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 6 min and presented a student showing prototypical and clearly substandard behavior with respect to communication and interpersonal skills. This DVD was considered to present a non-complex rating task. DVD 2—angina pectoris-lasted about 18 min and was considered to present a complex rating task with the student showing more complex behaviors 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 (Figs. 1, 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. 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).

[1] Verbal protocols refer to the collection of participants' verbalizations of their thoughts and behaviors, 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.

Overall performance

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

Fig. 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	

Fig. 2 6-Dimensional global rating scale clinical competencies (R2)

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

Table 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

single thought, idea or statement (see Table 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 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, ME). 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.

Results

Table 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 behavior (DVD 1). However, when observing the more complex behavioral 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 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' ($R1 \leq 2$). 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 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 behavior ($U = 62.5$, $p = .005$, $ES = 0.48$), whereas non-experts provided more descriptions ($U = 68.5$, $p = .009$, $ES = 0.45$). The

Table 2 Time needed for problem representation (T1) and performance ratings per DVD, for each group of raters

Variable	DVD 1 (prototypical, derma case)		DVD 2 (complex, cardio case)	
	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)	1390 (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]

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, $p < .05$)

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 behaviors ($U = 73$, $p = .01$, $ES = 0.42$). Experienced raters also generated significantly more interpretations when filling out the six-dimensional global rating scale ($U = 63$, $p = .004$, $ES = 0.48$).

Table 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.

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 behavior, but our expectations are confirmed for the case with complex trainee behavior, 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 behavior, 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 behavior may be explained by the rating stimulus in our study. The dermatology case may have been too short, and the succession of typical student behaviors 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 behaviors, 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 behavior.

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 behaviors, whereas non-experienced raters provided more literal descriptions of the observed behavior. 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

Table 3 Percentages of statements in verbal protocols for experienced raters (Exp) and non-experienced raters (Non-Exp)

Variable	Overall (VP1 + VP2 + VP3 + VP4)		VP1		VP2		VP3		VP4	
	Exp *Mdn (IQR)*	Non-Exp *Mdn (IQR)*	Exp *Mdn (IQR)*	Non-Exp *Mdn (IQR)*	Exp *Mdn (IQR)*	Non-Exp *Mdn (IQR)*	Exp *Mdn (IQR)*	Non-Exp *Mdn (IQR)*	Exp *Mdn (IQR)*	Non-Exp *Mdn (IQR)*
Descriptives	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)

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

VP1 verbal protocol at T1, initial representation of student behavior; *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, $p < .05$]

judgments. They integrate relevant background information and observed behaviors 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 behavior 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 in 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 intelligence, 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 et al. 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 behavior 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 behavioral detail (Murphy and Balzer 1986; Lievens 2001), whereas non-experienced raters may be more accurate at the behavioral 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 behaviors 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 behaviors.

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 behaviors 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 behavior 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.

Acknowledgments The authors would like to thank all the general practitioners who participated in this study. The authors would also like to thank Susan van der Vleuten and Meike Elferink (ME) for their support and assistance in data gathering and data analysis; and Mereke Gorsira for critically reading and correcting the English manuscript.

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.

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., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science: A Multidisciplinary Journal, 5*(2), 121–152.

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

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 DI (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–S81.

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. *Annual Review of Psychology, 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). New York: 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: Sage Publications Ltd.

Gray, J. D. (1996). Global rating scales in residency education. *Academic Medicine, 71*, S55–S61.

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. 205–230). 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*, 471–476.

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*, 880–889.

Norman, G. R., Brooks, L. R., & Allen, S. W. (1989). Recall by expert medical practitioners and novices as a record of processing attention. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 15*, 1166–1174.

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.

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.

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

Van De Wiel, 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.

Van der Vleuten, 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. 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.

Williams, R. G., & Dunnington, G. L. (2004). Prognostic value of resident clinical performance ratings. *Journal of the American College of Surgeons, 199*, 620–627.

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), 949–957.