Opening the black box of clinical skills assessment via observation: a conceptual model

Jennifer R Kogan,[1] Lisa Conforti,[2] Elizabeth Bernabeo,[2] William Iobst[2] & Eric Holmboe[2]

OBJECTIVES This study was intended to develop a conceptual framework of the factors impacting on faculty members' judgements and ratings of resident doctors (residents) after direct observation with patients.

METHODS In 2009, 44 general internal medicine faculty members responsible for out-patient resident teaching in 16 internal medicine residency programmes in a large urban area in the eastern USA watched four videotaped scenarios and two live scenarios of standardised residents engaged in clinical encounters with standardised patients. After each, faculty members rated the resident using a mini-clinical evaluation exercise and were individually interviewed using a semi-structured interview. Interviews were videotaped, transcribed and analysed using grounded theory methods.

RESULTS Four primary themes that provide insights into the variability of faculty assessments of residents' performance were identified: (i) the frames of reference used by faculty members when translating observations into judgements and ratings are variable; (ii) high levels of inference are used during the direct observation process; (iii) the methods by which judgements are synthesised into numerical ratings are variable, and (iv) factors external to resident performance influence ratings. From these themes, a conceptual model was developed to describe the process of observation, interpretation, synthesis and rating.

CONCLUSIONS It is likely that multiple factors account for the variability in faculty ratings of residents. Understanding these factors informs potential new approaches to faculty development to improve the accuracy, reliability and utility of clinical skills assessment.

*Medical Education 2011: **45**: 1048–1060*
doi: 10.1111/j.1365-2923.2011.04025.x

[1]Department of Medicine, University of Pennsylvania, School of Medicine, Philadelphia, Pennsylvania, USA
[2]American Board of Internal Medicine, Philadelphia, Pennsylvania, USA

Correspondence: Jennifer R Kogan, Associate Professor of Medicine, Director of Undergraduate Education, Department of Medicine, University of Pennsylvania, School of Medicine, 3701 Market Street, 6th Floor, Suite 640, Philadelphia, Pennsylvania 19104, USA.
Tel: 00 1 215 615 0503; Fax: 00 1 215 662 7979;
E-mail: jennifer.kogan@uphs.upenn.edu

INTRODUCTION

The assessment of the clinical skills required for patient care, such as history-taking, physical examination, counselling and interpersonal skills and professionalism, remains fundamental to the assessment of residents' clinical competence.[1–3] Accurate observations and high-quality feedback about clinical skills are requisite for the development of expertise[4] and faculty members are expected to observe and provide feedback to residents about their clinical skills.[5,6] Additionally, in response to calls to limit high-stakes final examinations, greater emphasis is now placed on the continuous assessment of skills in the clinical workplace[7–11] and such assessment must be accurate and valid.

Tools to facilitate the direct observation of residents' clinical skills with patients have been developed and published.[12] However, clinical skills performance ratings are subject to many sources of rating error.[13–15] Although many studies describe the poor accuracy and reliability of performance ratings, few have explored the factors underlying this rater variability. Raters themselves explain the largest component of variance in ratings,[16] and poor inter-rater agreement may result from differences in observer gender, ethnicity, experience or clinical competence.[17–20] However, these studies have largely focused on raters' traits and characteristics (most of which are immutable or unchangeable) that impact ratings without providing insight into *why* raters rate as they do or why these traits are associated with differences in rating behaviours.

Despite the numerous tools available to assess residents' encounters with patients, information regarding best practices in how to train raters to use them is relatively scarce. Approaches to minimise rating error have been described,[14] but the effectiveness of faculty development has been variable.[21,22] We postulate that there are likely to be other factors, still undefined, that may explain the poor reproducibility and inaccuracy of clinical ratings. An improved understanding of these factors could potentially inform more effective approaches to faculty development and thereby move direct observation forward as a keystone of assessment in competency-based medical education.[23,24] The purpose of the current study, therefore, was to explore, using qualitative methodology, factors that impact faculty assessment of residents, specifically in terms of how they judge and rate residents after observing their clinical skills with patients. We have previously reported the quantitative results of this study, which focused on the relationship between faculty members' demographics and clinical skills and their rating behaviours.[20]

METHODS

Sample

Programme directors from seven university-based and nine community-based, university-affiliated internal medicine residency programmes in a large urban area in the eastern USA were e-mailed and asked to identify general internal medicine out-patient faculty resident preceptors potentially interested in participating in a study about resident assessment. In the USA, the internal medicine residency refers to the 3-year training period that follows 4 years of medical school. A total of 114 faculty staff were subsequently e-mailed and invited to participate; recruitment stopped after the first 48 faculty members replied based on an *a priori* power calculation for the quantitative component of this study.[20] Table 1 describes additional sample characteristics. Of the 48 faculty staff who agreed to participate, 44 (92%) completed the study; four faculty members dropped out (for reasons of personal conflict, family illness or lack of hospital coverage).

Study design and data collection

Data collection occurred between March and August 2009 with three to six faculty staff participating each study day. Prior to their assigned study day, faculty members completed a web-based demographic questionnaire that has been previously described.[20] On their study day, faculty members individually watched four videos and two live scenarios of a standardised postgraduate year 2 (PGY2) resident (SR) taking a history, performing a physical examination or counselling a standardised patient (SP).[20] The live cases also scripted resident receptiveness to feedback. These cases were previously used with medical residents and each case was scripted to depict a PGY2 resident whose performance was unsatisfactory, satisfactory or superior for content (history taking, examination, counselling) and interpersonal skills (some cases portrayed superior content but unsatisfactory interpersonal skills). Initial error scripting (by JRK) was based on actual resident performance norms. The study team reviewed scripts to confirm that they reflected predetermined performance levels. For the video cases, volunteer medical residents trained on a single script, practised

Table 1 Demographics of participants (n = 44) in a qualitative study of direct observation of clinical skills, 2009

Characteristic	
Age, years, mean (SD)	44.2 (8.7)
Male, *n* (%)	25 (57)
Rank, *n* (%)*	
Instructor	4 (9)
Assistant professor	19 (43)
Associate professor	15 (34)
Professor	4 (9)
Affiliation, *n* (%)	
Community-based	20 (46)
University-based	24 (54)
Out-patient precepting experience, years, mean (SD)	12.4 (7.5)
Non-precepting out-patient clinical work, %, mean (SD)	46.2 (25.0)
Prior participation in workshop on assessment of residents in a clinical setting, *n* (%)	20 (44)
Prior participation in workshop on giving feedback, *n* (%)	23 (52)
Use of mini-CEX in past year to assess residents, *n* (%)	39 (89)

* Two participants did not report their rank
Mini-CEX = mini clinical evaluation exercise

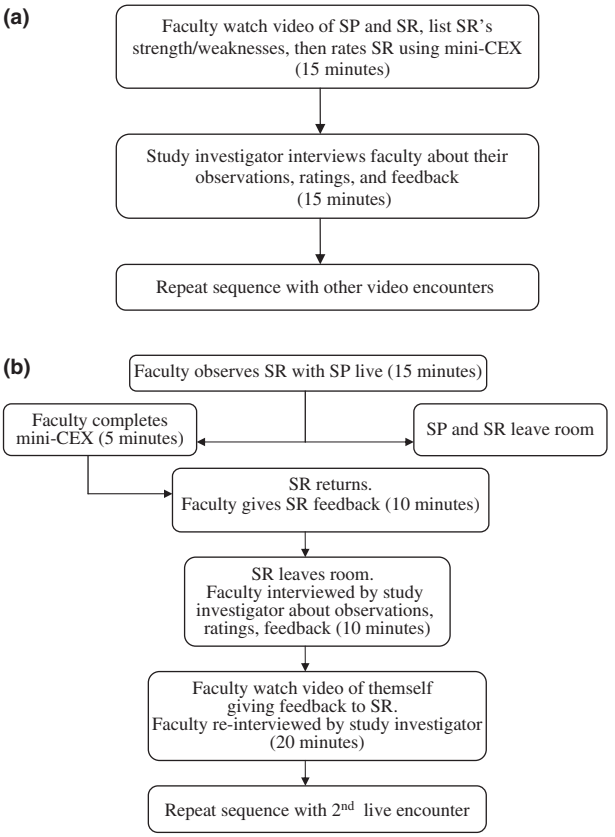

Figure 1 Study protocol for (a) video encounters and (b) live encounters between standardised residents and patients. SP = standardised patient; SR = standardised resident; mini-CEX = mini clinical evaluation exercise

with the SP and were videotaped once their performance accurately represented the intended performance level. For the live cases, residents were given scripts to guide their performance and receptiveness to feedback.

After watching each of four video encounters (Fig. 1a), faculty staff completed a mini-clinical evaluation exercise (mini-CEX). The mini-CEX, developed by the American Board of Internal Medicine (ABIM) to provide residents with feedback about their history-taking, physical examination, counselling and interpersonal skills, details seven competencies that are rated on a 9-point scale (1–3 = unsatisfactory, 4–6 = satisfactory, 7–9 = superior).[25,26] Faculty members were then interviewed individually for 15 minutes by a trained study investigator using a semi-structured interview guide. Appendix S1 (online only) presents examples of primary and secondary interview questions. Videos were shown in a random order and each faculty member was interviewed by at least three interviewers. Following the video scenarios, faculty members observed two live representations of an SR taking a history, conducting an examination and counselling an SP (Fig. 1b). Following each encounter, faculty staff rated the SR using the mini-CEX and provided the SR with up to 10 minutes of feedback, which was video-recorded. Faculty members were then interviewed individually by a study investigator for 30 minutes using the semi-structured interview (Appendix S1). Faculty members were asked about the feedback encounter before and after watching a DVD of themselves giving feedback to the SR. All interviews were video-recorded and transcribed verbatim with identifying information about the participants removed, and all transcripts were reviewed for accuracy. The University of Pennsylvania School of Medicine Institutional Review Board approved the study. The work was carried out in accordance with the Declaration of Helsinki, including, but not limited to, a proviso that no potential harm to participants could occur. The anonymity of participants was guaranteed and all participants provided informed consent.

Data analysis

We utilised a grounded theory approach to analyse the data for emergent themes and to develop a thematic coding structure.[27] We selected grounded theory because little is known about the observation and evaluation process and we wished to avoid restricting ourselves to current hypotheses or inferences from prior studies.[28] Transcripts were sampled for coding across faculty participants, SP cases and interviewers.[27] Two researchers (JRK, LC) independently coded and used constant comparative techniques to develop a preliminary coding structure.[27] A portion of the transcripts were also coded by additional study team members (EB, WI, EH) to review, further define and refine the coding structure. Refinement of the coding structure continued as analysis progressed. Coding was terminated when theoretical saturation was achieved and when all team members agreed upon the final interpretation of the data. In total, 56 of 172 video interviews (33%) and 29 of 88 live interviews (33%) were coded. NVivo Version 2.0 (QSR International Pty Ltd, Melbourne, Vic, Australia) was used to organise and analyse the coding structure.

RESULTS

We identified four themes that help to explain the variability in faculty judgements and ratings of SRs. These themes include: (i) the use of variable frames of reference during observation and rating; (ii) the role of inference; (iii) the use of variable approaches to synthesising judgements into numerical ratings, and (iv) factors external to resident performance that influence ratings.

Theme 1. Frames of reference during observation and rating

Faculty members drew from a number of frames of reference (i.e. standards for judgement or comparison) when observing a resident and judging and rating his or her performance (interpreting the observation). These various frames of reference enabled the comparison of the resident's performance with: (i) performance by oneself; (ii) the performance of other doctors (both residents and practising doctors), and (iii) a standard of performance considered to be necessary for patient care.

Using self as a reference

Many faculty staff used themselves as a frame of reference when making judgements and assigning ratings. Most frequently, faculty members compared resident performance with how they perceived themselves to practise:

'He walked in and he's like, I have some bad news. I would never do that.' (Faculty member C3, video case 2)

Faculty members' perceptions of their own clinical strengths or limitations at times mediated their judgements and ratings, as well as their comfort with the encounter. As one faculty member stated:

'A truly seasoned clinician would say, ''Is that the only thing on your mind? Is there something more on your mind?'' And the resident certainly did not do that. And hopefully, I would. I keep thinking – that's always what I think when I watch this: Am I doing this? Would I expect this of myself? Hence is it fair to expect this of someone junior to me?' (Faculty member I3, video case 3)

Competencies believed by faculty members to be especially important and which were prioritised in feedback also framed ratings:

'The first thing I always pick is the interpersonal communication portion because it just happens to be 90% of what our job is. So no matter what the science or disease is – it very much comes down to how you relate to the patient and what kind of rapport you can set up with the patient. So it's always going to be [part] of what I discuss. Moreover it's the thing that is the hardest to get across, it's the hardest skill to learn.' (Faculty member M1, video case 1)

Faculty staff also referred to comparisons of resident performance with the faculty member's perception of his or her own performance as a resident. Faculty members also framed ratings based on how they would want to receive care as a patient:

'A lot of it is just instinct. A lot of it is when I've been a patient myself what I've looked for in a good doctor.' (Faculty member I4, video case 1)

Using other doctors as a reference

Many faculty members compared resident performance with that of residents at a similar stage:

'For her level of training, she's a PGY2, I felt like this was actually better than [the] average [performance] I would expect with PGY2. If this was a PGY3, I probably would have given slightly lower ratings, even

though I felt like a lot of things were good. I expect that by PGY3 they should do something a little better than this.' (Faculty member F1, video case 4)

However, some faculty staff questioned whether ratings should be based on the resident's PGY level:

'I'm probably less likely to give them a 4 if they're a PGY1 and this is their third clinic. I feel like they're more probably satisfactory for where they are and what they should know. But it's a question. Should we be giving them 3s based on they are unsatisfactory and that's okay because they've only just been doing this? So it's not like they're going to fail? But if they're a PGY3 and they're getting 3s, then that is a big deal. So I think there's no standardisation as far as I can tell, as far as the way we're trained to do this.' (Faculty member A7, video case 4)

Some faculty staff also compared resident performance with that of practising doctors. Many faculty members acknowledged that some practising doctors have deficient clinical skills and this led them to question what it might be reasonable to expect of a resident:

'Having been in practice, I have met people like him [the resident] and so the question is, what are our expectations? And what is realistic in terms of expectations? It's very pessimistic. I realise as I am saying this. I feel that I encounter more and more doctors who are cynical and who are less interested in developing a patient-centred relationship with their patient.' (Faculty member A3, video case 2)

Using patient outcomes as a reference

A few faculty staff used patient outcomes (e.g. achieving a correct diagnosis, the likelihood that the patient would return for follow-up care, patient compliance with medication) as a frame of reference for ratings:

'Some of it is not so much [about] what I want the resident to get skill-wise, but we have to make sure that this patient is safe in his surroundings, meaning other patients are safe too. So the patient care issue needs to be really, really high.' (Faculty member I2, video case 3)

Additional frames of reference

A few faculty members used existing frameworks to guide them in assessing residents. For example, this faculty member assessed the resident's history-taking skills in a patient with possible depression:

'In terms of what would have made it a 6? …she asked most of the questions of SIGECAPS [Sleep, Interest, Guilt, Energy, Concentration, Appetite, Psychomotor, Suicidal]. One of the things that's often left out by residents is to ask about suicidal ideation because it's an awkward thing.' (Faculty member M1, video case 4)

For others, articulating the standard for evaluation was difficult. Instead, some faculty staff referred to having a 'gut' feeling that drove evaluation. Others had difficulty in verbalising how they moved from observations to judgement and commented that the transition represented a gestalt or simply that they were uncertain of which framework they were making judgements and ratings against. Some faculty staff found the assessment of interpersonal skills particularly challenging because these skills were felt to be more subjective and difficult to quantify:

'It's the more nebulous, less quantifiable skills, such as connecting to the patient, which 10 people can do in 10 different ways. That's the problem. But she's not doing it in any one of those 10 ways. The nebulousness of interpersonal warmth and communication is so hard for me to relay to a resident … because I don't feel there's any one absolute way to do it.' (Faculty member I3, video case 1)

Importantly, our data show that the ways in which faculty staff implemented these frames of reference were complex, dynamic and highly variable. Many faculty members shifted between frames of reference both within and between encounters:

'So I think just using my gut, observing things … I think for the positives it's probably a little bit of, you've seen so many residents over the years, and there are things that just stand out as being like, huh, that took me a while to figure out how to do that, and they do that really well. So, for example, for him, he walked in, he got a chief complaint immediately and very easily asked a very nice open-ended question to elicit the information… And you could tell it was a skill. And it's a skill that not everyone – because you watch residents a lot – a lot of them don't have that. He was checking information with her all the time, and that is something that took me a while to learn. So that is, I think, the criteria I was using to judge something that's being done well.' (Faculty member C6, live case 2)

Theme 2. The role of inference

We found that inferences about residents and their performance were prominent during assessment. Faculty members used concrete data (resident actions), selected from those actions (consciously or subconsciously), affixed meaning and interpretation to those actions and made assumptions from which they frequently drew conclusions. Often, inferences were of the 'high' level, meaning there was significant interpretation based on the behaviour witnessed. Table 2 provides examples of how the same behaviour was interpreted differently by different faculty staff watching the same video. Inferences were made about the residents' feelings (i.e. their levels of confidence and comfort), personalities, skills (i.e. knowledge base and potential), motivation to improve, and prior experiences and preparation.

A few faculty staff seemed to be aware that they made inferences, as illustrated by this doctor:

'The one thing that struck me through this entire visit was he [the resident] had his arms kind of crossed. That could mean different things to different people in terms of body language. It means I'm either closed to you, or I'm very comfortable with this, but it seems less likely. It could also be [representative of] an uncomfortable feeling on the part of the resident. He wants to act like he's comfortable, but internally, he's very anxious about breaking bad news, so there's a number of ways you can look at that…' (Faculty member L2, video case 2)

However, many faculty members failed to recognise when they made subjective inferences and consequently made numerous assumptions about residents' performance.

Theme 3. Variable approaches to synthesising judgements to numerical ratings

Our data showed significant variability and uncertainty surrounding how to translate a judgement about the resident into a numerical rating, especially the overall mini-CEX rating. However, a few strategies emerged. Some faculty members chose to average all of the individual mini-CEX competencies:

'So I believe what I first do is I just sort of do the numbers in my head. You know quickly just boom, boom, boom what's the average… I just add them up and do them mentally in my head.' (Faculty member J2, video case 3)

Others used non-compensatory grading:

Interviewer: 'Did the humanism pull her up?'

Faculty member: 'I don't think so, because competence is still, you know, absent competence – a humanistic physician who's not competent is a very dangerous person.' (Faculty member F2, live case 2)

Some faculty staff weighted ratings according to the encounter's focus or purpose:

'I would think to myself, what was this case about? And so I kind of weigh that qualitatively… If the case was revolving more around counselling, then I might, not discount, but lower the weight on the interviewing, organisation and stuff like that and really focus on how they did in the counselling.' (Faculty member D3, video case 2)

Many faculty staff struggled to translate their judgements to a numerical rating. Faculty members described their own lack of understanding about the meaning of the numbers, their inability to discriminate along a 9-point scale, and their uncertainty regarding how to synthesise ratings:

'I mean there is a scale from 1 to 9, but there's no guidelines on which one is what. And even if there is, there is always greyness in it.' (Faculty member M2, live case 1)

'I tell residents that… I can't make nine divisions. I can make three divisions in how I see them function… I think an unsatisfactory, maybe two satisfactories and then a superior is about the best I can do.' (Faculty member B1, video case 4)

'One of the things that I still struggle with [about] this 9-point scale is that there's no general rule… Is this sort of an average of all the things they've done, or do you have to get a 4 on everything in order to get a 4? I don't even think I have my own rule on that…' (Faculty member C6, live case 2)

Theme 4. Factors external to resident performance that drive ratings

Several additional factors influenced faculty staff ratings, including context (the complexity of the encounter, the resident's prior experience, the faculty–resident relationship) and response to feedback (by the resident, by the faculty member, by the institution). Some of these factors may help to explain common rating errors such as range

Table 2 Examples of inference during observation and assessment of standardised resident performance in video encounters

Video	Faculty member's assessment	Inference
Male patient with acute dysuria	'He's a shy guy… who I think may have had [a] distant accent. Like he was accented when he was a kid, so maybe he grew up in a culture [in which] sex was not an appropriate topic of conversation, even with a physician… or maybe he grew up in a religious background where that's different. And having a little bit of background on that would be helpful, for me, because this guy is like amazing and he's just like [a] hard worker, super nice…' (Faculty member B2, video case 3)	Personality (shy, nice) Culture Work ethic Competence
	'I thought maybe he was just a little unaccustomed to the situation' (Faculty member L1, video case 3)	Familiarity with scenario
	'Yeah, he wasn't shy about addressing using barrier techniques and using condoms, and he seemed very comfortable addressing everything that needed to be addressed' (Faculty member D4, video case 3)	Comfort
	'I think I observed two stiff people. Both patient and resident. I've seen worse, but still, they seemed both a little uncomfortable and embarrassed, perhaps, with the topic being discussed' (Faculty member I3, video case 3)	Comfort
Delivering a new cancer diagnosis	'He knew the answers to some things that she wanted to know… Obviously he shouldn't try and give her some big snow job at the time and bombard her with data and statistics and all. But she wanted to know, what can I anticipate? …and, I think, he wilfully withheld things from her that he could've told her that might've been helpful for her… He seemed to dump… He very much took the view: ''This is not my problem.'' And he was clearly organised and efficient; he wanted to get this over with' (Faculty member F2, video case 2)	Intentions Ownership
	'But first and foremost, his body language told this patient that he couldn't wait to get out of the room. He stood. He had his arms folded. He was clearly uncomfortable… He was clearly uncomfortable with the scenario and he needs to get, you know, he needs to read how to give bad news, you know, he needs to learn that type of stuff…' Interviewer: 'How did you know he was uncomfortable?' Faculty member: 'You could see it, body language. He had his arms folded, he crossed his legs. He was tight like he was holding himself. He was very monotone' (Faculty member D3, video case 2)	Wishes Comfort Knowledge
	'It's kind of a tough situation to give bad news to a patient. Probably this resident, well, maybe this resident has not had a lot of experience with it. …I mean it's not like explaining discharge medications or something you do 10 times a day. I thought he did some things pretty well. And I got the impression that he would do better in future encounters, at least I hope so, because he, he seemed kind of ill at ease, and I perhaps mistakenly attributed that – I don't know if it was right or wrong, but attributed it to sort of maybe being not real familiar with how to do this really well. I got the impression that he had thought about how to give bad news because he did do some of the things that we advise trainees to do when they're giving bad news and then once he got into the situation more he seemed a little bit less comfortable and not quite knowing what to do' (Faculty member L1, video case 2)	Prior experience Comfort

restriction (avoiding ratings at the upper [severity error] and lower [leniency error] ends of the scale) and the halo effect.

Context

Contextual factors such as the complexity of the clinical scenario and perceptions of the resident's familiarity with a clinical situation influenced how faculty members translated their observations into ratings:

'Sometimes it's difficult talking to patients about sexual histories and STDs [sexually transmitted diseases], so it was a difficult counselling session. And who knows how many of these he's done before. The way that I think about it is if you perform pretty darn good in a really hard situation, that's a lot better than performing pretty darn good in a pretty easy situation.' (Faculty member L1, video case 3)

The duration of the resident–faculty relationship also impacted ratings. Faculty members, referring to their experiences with actual residents, explained that when they had a longitudinal relationship with a resident, they knew what that resident had already received feedback on. The repetition of a mistake by the resident after feedback resulted in rater stringency:

'If I could remember that this is a resident I had already talked [to] about how to take a sexual history and he did this kind of job, I would probably rank him a little bit lower, saying [that] we have talked about this.' (Faculty member A1, video case 3)

By contrast, a pre-existing positive relationship with a resident was sometimes associated with rater leniency and the halo effect:

'If you have a relationship with a trainee, student or resident, it's hard not to have that impact someone. So if it's a resident specifically, who you really like, you're probably more likely to cut them slack, as opposed to a resident who you either haven't worked with or by reputation is very biomedical, kind of no interpersonal qualities… So I think that by knowing someone and having a relationship… maybe it's that you're willing to overlook the little things.' (Faculty member C2, video case 1)

Response to feedback

Our data showed that faculty members' inferences about residents' responses to feedback influenced their ratings. Emotions frequently stemmed from concern about residents' reactions to numerical ratings that might be either high ('How will they grow and get better?' [Faculty member B4, video case 4]) or low:

'If anyone's in the satisfactory category, I tend to put them in the 6 range because I don't want to be having a conversation about why it wasn't a 6 instead of a 5. But [what] I really want to say to this resident is, you were fine, you were at your level of training, that's where I want you to be. I don't want to be negotiating about why I picked it as a 4 or 5 or 6 in that range. So, I pick it as a 6 so that takes that conversation off the table… I have dealt with enough students and residents that [I know] I don't want people to focus so much on the number. I really want them to focus on here's what you did well, here's what you might do differently.' (Faculty member A1, video case 3)

In addition, the faculty member's own emotional response to providing constructive feedback (i.e. feeling mean or unkind; being 'demoralising') and concern about the emotional impact on the resident and how the resident might perceive the faculty member seemed to mediate assessment:

'People don't like to give low scores because it doesn't show well for them… And the answer that I've gotten, and I haven't been very satisfied with it, is that you can be a popular doctor or a good, good doctor with residents and medical students.' (Faculty member A7, video case 4)

By contrast, other faculty staff were focused on their roles and responsibilities as coaches:

'It felt just fine [giving feedback] because that's my job. It's to make them the best doctors they can be. I'm their coach. This is all done in the spirit of making them the best team players they can be… It's the most time-consuming, but it's also the most rewarding when you take those kids that are down there and bring them up here. So I have no problem, whether I have to give good feedback, or bad feedback.' (Faculty member C1, live case 2)

Finally, several faculty members described the role of the broader institutional culture in guiding their ratings:

'I do a lot of CEXs and it's uncomfortable to give a resident a 4 or below. I feel like I can express my dissatisfaction well in my comments without having to negotiate whether this was a 4 or 5 for the resident…

I think the value of this is the comments. [This attitude] came from dealing with residents or medical students and having to defend the grades I gave to the clerkship directors, to the deans. Sitting in very uncomfortable meetings with the dean and the student who I graded very poorly and having to sit there and defend my score. And I remember – and this was early on, when I was a faculty member – thinking like, this is silly; I don't want to be in this situation again.' (Faculty member A1, video case 2)

DISCUSSION

Using a grounded theory approach, we identified several factors influencing how faculty members judge and rate residents during clinical skills assessment, including the use of variable frames of reference, the use of high degrees of inference, the use of variable approaches to synthesising observations and judgements into numerical ratings, and factors that are external to resident performance. These themes highlight the variability in the entire process of direct observation, judgement and ratings.

From these emerging themes, we have derived a model describing the process of direct observation of clinical skills and the factors affecting observations, judgements and ratings (Fig. 2). Faculty staff bring to resident clinical skills assessment an amalgam of characteristics that potentially impact on observations and assessment. These characteristics include age, gender, clinical and teaching experience, clinical and educational competence, and attitudes and emotions related to observation and feedback. Faculty members observe resident and patient interactions through two lenses, one of which concerns a frame of reference (whereby faculty members use their own or other doctors' performance, or patient outcomes, as a yardstick against which to compare resident performance) and one of which refers to inference which further shapes the meaning and interpretation assigned to observations. The faculty member's observation of the trainee with the patient occurs within and is influenced by contextual factors including the clinical system (i.e. familiarity with the patient, patient complexity, organisation of the clinical unit, etc.) and educational system (i.e. institutional culture and oversight). During and after observation, the faculty member interprets and synthesises his or her observations into a rating. However, the process is not neat, predictable or straightforward. Multiple additional influences that can impact on ratings include anticipated feedback, institutional culture and encounter complexity. These influences further support the importance of context in observation, feedback and ratings.[29,30] Interpretation and synthesis of observations, ratings and feedback become experiential modifiers that subsequently impact on the faculty member and the lens through which he or she makes observations, both within a single encounter as well as in future encounters.

These complex interactions can be supported by situated cognition theory which contends that an individual's thinking, knowing and processing are uniquely tied to and inextricably situated within (and

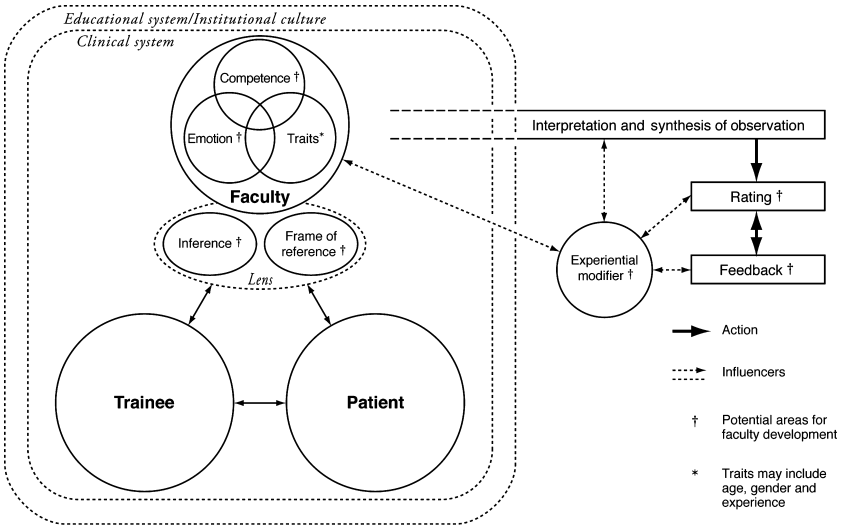

Figure 2 Conceptual model: process of direct observation of clinical skills and factors affecting observations, judgements and ratings

cannot be completely separated from) the specific social situations within which those thoughts and actions occur.[29,31,32] Situated cognition contends that failing to acknowledge the contributions of the setting leads to a perspective on thinking that cannot fully capture the construct;[33] that is, thinking and acting are context-specific to the environment, which plays a role equal to that of the person.[31,33] Using situated cognition as a theoretical framework helps to provide a framework for our findings regarding the observation and assessment of trainees' skills in the clinical setting. We have found that factors including the trainees, the clinical and educational setting, the institutional culture and the faculty members themselves all become important and that they interact in a myriad of dynamic and unique ways. In terms of progressing, it will be important to understand these dynamic interplays and to acknowledge that some of the factors that affect observation and feedback are immutable (faculty age, gender), whereas many are potentially modifiable and could be addressed via faculty development.

Our data suggest that faculty staff approach assessment using multiple frames of reference. Poor inter-rater reliability of clinical skills assessments can be explained if one faculty member rates performance based on PGY level, another uses a standard of self, and another makes a rating based on a gestalt. We were struck by how often self was used as a frame of reference, particularly by comparing a resident's performance with one's current practice style. This finding has important implications because faculty staff have variable clinical skill proficiency.[20,34–36] Connecting competency frameworks with the work environment of patient care and entrustable professional activities is important.[11] Yet, in the present study, faculty staff rarely used evidence-based frameworks describing best clinical skills practices (e.g. informed decision making, patient communication) or patient outcomes[37–39] to anchor observation and assessment. Variable frames of reference are also problematic for residents. If the standard used to assign ratings is not articulated to the resident, the resident lacks a context in which to interpret his or her assessment. Furthermore, variable assessments of the same performance can potentially undermine the feedback process if residents preferentially dismiss or censor constructive feedback that is incongruent with their self-image.[40]

High-level inference has the potential to undermine feedback quality because feedback is potentially based on faulty assumptions. Reaching conclusions about performance requires faculty staff to use real data (resident behaviours), select from those behaviours, and affix meaning to them. These assumptions form conclusions which, in turn, lead to actions (the rating and feedback).[41] Particularly striking was the obvious use of inference yet the relative infrequency of questioning and active testing of the inferential assumptions with the resident. Although the faculty member did not have the opportunity to talk with the SR in the video encounters, similar inferences were made and not questioned during the live observations when the faculty member met with the SR for feedback. Faculty staff should be alerted to the prevalence of inference; training faculty to 'test' their assumptions during feedback (e.g. by asking: 'I had the feeling from watching you that you were pretty uncomfortable with this case; what are your thoughts on this?') may be valuable.

We found that faculty members struggled to translate their observations and judgements into numerical ratings. Overall impressions of resident performance do not represent a simple linear addition of the various dimensions being assessed and the weighting of dimensions does not necessarily improve a faculty member's sense of a resident's competency.[42,43] Poor inter-rater reliability of quantitative ratings can be addressed by ensuring that multiple observations of a resident by multiple raters at multiple time-points occur.[44] However, faculty members who use rating forms need to know what the numbers mean and how to select an overall rating. To our knowledge, faculty development has not addressed how observations and assessments should be synthesised into an overall rating. Our results also raise questions about whether we should rely more on comments than on numerical ratings in resident assessment. At the resident level, numbers may become less important and may be potentially counterproductive if faculty staff are overly affected by the challenges of assigning ratings. A particularly important theme that we identified is how a faculty member's anticipation of impending feedback impacts the rating process. Moving from observation to judgement to rating to feedback is a complex and interdependent process in which expectations of how feedback will play out, for both the resident and the faculty member, influence faculty ratings, introducing further variability.

We believe that awareness of the aforementioned themes should inform and enhance faculty development in clinical skills assessment. Our findings suggest that there is a need to ensure that faculty staff approach assessment with a shared standard or mental model, ideally shifting from a self-based to a

criterion-based framework. Knowledge of expected competencies and elucidation of milestones at particular levels of training could be valuable to faculty staff who are required to make assessments.[45] Identifying and reviewing available evidence-based clinical skills frameworks may enable faculty members to rely less on their own practice style in making assessments. Faculty development should recognise and address concerns about the feedback process and its impact on ratings. Identifying ways that faculty staff might better embrace assessment as a necessary component of deliberate practice and development of expertise could be important.

Although we sampled across community- and university-based programmes and reached thematic saturation, we focused only on general internal medicine faculty staff in one region of the USA, who largely precept in the ambulatory setting. Additional work is needed to determine the reproducibility of findings across disciplines and settings, including situations in which faculty staff interact with their own residents during clinical encounters with real patients. Further work is also needed to explore the interactions between the individual factors we have identified.[29]

We have attempted to explore factors that may impede the ability of the medical education community to achieve accurate, reproducible assessment or evaluation in a competency-based approach to training. Our findings have prompted the development of our new conceptual framework for understanding how faculty staff translate observations of trainee performance with patients into judgements and ratings. Further investigation is needed to evaluate whether new faculty development approaches can arm faculty members with the skills necessary to improve the accuracy and reliability of assessment using direct observation.

Contributors: JRK and LC made substantial contributions to the study conception and design, the acquisition, analysis and interpretation of data, and the drafting of the article. EB, WI and EH made substantial contributions to the study conception and design, and the acquisition, analysis and interpretation of data. All authors contributed to the critical revision of the article and approved the final manuscript for publication.

Acknowledgements: the authors wish to thank participating faculty members, the Drexel University College of Medicine's Standardised Patient Programme, the University of Pennsylvania School of Medicine's Patient Programme and the Penn Medicine Clinical Simulation Center, and the standardised patients and residents. We also thank the individuals who assisted in faculty interviews, including Rebecca Baranowski MEd, Benjamin Chesluk PhD, Steven Durning MD, Brian Hess PhD, Krista Hirschmann PhD, Lorna Lynn MD and Michael Pistoria DO, and Siddharta Reddy MPH for graphical assistance, and Steve Durning MD and Karen Hauer MD for their thoughtful review of the manuscript.

Funding: this study was funded by the American Board of Internal Medicine, from which JRK receives salary support.

Conflicts of interest: EH is in receipt of royalties from Mosby-Elsevier for a textbook on the assessment of clinical competence.

Ethical approval: this study was approved by the University of Pennsylvania School of Medicine Institutional Review Board.

REFERENCES

1 Kalet AL, Gillespie CC, Schwarz MD *et al.* New measures to establish the evidence base for medical education: identifying educationally sensitive patient outcomes. *Acad Med* 2010;**85** (5):844–51.
2 Accreditation Council for Graduate Medical Education. ACGME Outcomes Project. http://www.acgme.org/Outcome. [Accessed 24 October 2010.]
3 Frank JR. *The CanMEDS 2005 Physician Competency Framework.* Ottawa, ON: Royal College of Physicians and Surgeons of Canada 2005.
4 Ericson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Acad Med* 2004;**79** (10 Suppl):70–81.
5 Accreditation Council for Graduate Medical Education. ACGME Program Requirements for Graduate Medical Education in Internal Medicine. http://www.acgme.org/acWebsite/downloads/RRC_progReq/140_internal_medicine_07012009.pdf. [Accessed 24 October 2010.]
6 Training and Assessment. The Foundation Programme. http://www.gmc-uk.org/New_Doctor09_FINAL.pdf_27493417.pdf_39279971.pdf. [Accessed 24 October 2010.]
7 Miller G. The assessment of clinical skills/competence/performance. *Acad Med* 1990;**65** (9 Suppl):S63–7.
8 van der Vleuten C, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ* 2005;**39** (3):309–17.
9 Southgate L, Cox J, David T *et al.* The General Medical Council's performance procedures: peer review of performance in the workplace. *Med Educ* 2001;**35** (1 Suppl):9–19.
10 Hodges BD. A tea-steeping or i-Doc model for medical education? *Acad Med* 2010;**85** (9 Suppl):34–44.
11 ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med* 2007;**82** (6):542–7.
12 Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA* 2009;**302** (12):1316–26.

13 Williams RG, Klamen DA, McGaghie WC. Cognitive, social, and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003;**15** (4):270–92.
14 Woehr DJ, Huffcutt AI. Rater training for performance appraisal: a quantitative review. *J Occup Organ Psychol* 1994;**67** (3):189–205.
15 Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ* 2004;**38** (3):327–33.
16 Downing SM. Threats to the validity of clinical teaching assessments: what about rater error? *Med Educ* 2005;**39** (4):353–5.
17 McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk–dove effect') in the MRCP (UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Med Educ* 2006;**6**:42.
18 Carline JD, Paauw DS, Thiede KW, Ramsey PG. Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. *J Gen Intern Med* 1992;**7** (5):506–10.
19 Weingarten MA, Polliack MR, Tabenkin H, Kahan E. Variations among examiners in family medicine residency board oral examinations. *Med Educ* 2000;**34** (1):13–7.
20 Kogan JR, Hess BJ, Conforti LN, Holmboe ES. What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. *Acad Med* 2010;**85** (Suppl):25–8.
21 Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence. A randomised trial. *Ann Intern Med* 2004;**140** (11):874–81.
22 Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomised controlled trial. *J Gen Intern Med* 2009;**24** (1):74–9.
23 Shumway JM, Harden RM, Association for Medical Education in Europe. AMEE guide no. 25: the assessment of learning outcomes for the competent and reflective physician. *Med Teach* 2003;**25** (6):569–84.
24 Crossley J, Humphris G, Jolly B. Assessing health professionals. *Med Educ* 2002;**36** (9):800–4.
25 Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med* 1995;**123**:795–9.
26 Durning SJ, Cation LJ, Markert RJ, Pangaro LN. Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. *Acad Med* 2002;**77**:900–4.
27 Strauss A, Corbin J. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Newbury Park, CA: Sage Publications 1998.
28 Patton MQ. *Qualitative Research and Evaluation Methods*. Thousand Oaks, CA: Sage Publications 2001.
29 Durning SJ, Artino AR, Pangaro LN, van der Vleuten C, Schuwirth L. Redefining context in the clinical encounter: implications for research and training in medical education. *Acad Med* 2010;**85**:894–901.
30 Koens F, Mann KV, Custers EJFM, ten Cate OTJ. Analysing the concept of context in medical education. *Med Educ* 2005;**39**:1243–9.
31 Bredo E. Reconstructing educational psychology: situated cognition and Deweyian pragmatism. *Educ Psychol* 1994;**29**:23–5.
32 Kirshner J, Whitson JA. *Situated Cognition: Social, Semiotic and Psychological Perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates 1997.
33 Durning SJ, Artino AR, Holmboe E, Beckman TJ, van der Vleuten C, Schuwirth L. Ageing and cognitive performance: challenges and implications for physicians practising in the 21st century. *J Cont Educ Health Prof* 2010;**30**:153–60.
34 Ramsey PG, Curtis JR, Paauw DS, Carline JD, Weinrich MD. History-taking and preventive medicine skills among primary care physicians: an assessment using standardised patients. *Am J Med* 1998;**104** (2):152–8.
35 Paauw DS, Weinrich MD, Curtis JR, Carline JD, Ramsey PG. Ability of primary care physicians to recognise physical findings associated with HIV infection. *JAMA* 1995;**274** (17):1380–2.
36 Vukanovick-Criley JM, Criley S, Warde CM, Boker JR, Guevara-Matheus L, Churchill WH, Nelson WP, Criley JM. Competency in cardiac examination skills in medical students, trainees, physicians and faculty: a multi-centre study. *Arch Intern Med* 2006;**166** (6):610–6.
37 Braddock CH III, Edwards KA, Hasenberg NM, Laidley TL, Levinson W. Informed decision making in out-patient practice: time to get back to basics. *JAMA* 1999;**282** (24):2313–20.
38 Makoul G. Essential elements of communication in medical encounters: the Kalamazoo consensus statement. *Acad Med* 2001;**76** (4):390–3.
39 Deber RB. Physicians in health care management: 7. The patient–physician partnership: changing roles and the desire for information. *CMAJ* 1994;**151** (2):171–6.
40 Swann WB Jr. Self-verification: bringing social reality into harmony with the self. In: Suls J, Greenwald AG, eds. *Social Psychological Perspectives on the Self*, Vol. **2**. Hillsdale, NJ: Lawrence Erlbaum Associates 1983;2.
41 Senge PM, Kleiner C, Roberts C, Ross R, Smith B. *The Fifth Discipline Fieldbook: Strategies for Building a Learning Organization*. New York, NY: Doubleday 1994.
42 Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr G. Toward authentic clinical evaluation: pitfalls in the pursuit of competency. *Acad Med* 2010;**85** (5):780–6.
43 Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Acad Med* 1999;**74**:1129–34.
44 Govaerts MJ, van der Vleuten CP, Schuwirth LW, Muijtjens AM. Broadening perspectives on clinical performance assessment: rethinking the nature of

in-training assessment. *Adv Health Sci Educ* 2007;**12** (2):239–60.
45 American Board of Internal Medicine. Milestones Framework. http://www.abim.org/milestones/public/ [Accessed 24 October 2010.]

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Examples of interview questions asked of internal medicine faculty members in the Philadelphia area in 2009.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than for missing material) should be directed to the corresponding author for the article.

Received 6 December 2010; editorial comments to authors 25 January 2011; accepted for publication 21 March 2011