Received: 1 September 2021 | Revised: 3 February 2022 | Accepted: 26 February 2022

DOI: 10.1111/medu.14789

RESEARCH ARTICLE

ASSESSMENT

Safeguarding fairness in assessments—How teachers develop joint practices

Linda Barman[1] | Cormac McGrath[2] | Staffan Josephsson[3] | Charlotte Silén[4] | Klara Bolander Laksov[2,4]

[1]Department of Learning in Engineering Sciences, KTH Royal Institute of Technology, Stockholm, Sweden

[2]Department of Education, Stockholm University, Stockholm, Sweden

[3]Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden

[4]Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden

Correspondence
Linda Barman, Department of Learning in Engineering Sciences, KTH Royal Institute of Technology, Osquars backe 31, Stockholm 100 44, Sweden.
Email: lbarman@kth.se

Funding information
Stockholm County Council, Grant/Award Number: SLL 20110191

Abstract

Introduction: In light of reforms demanding increased transparency of student performance assessments, this study offers an in-depth perspective of how teachers develop their assessment practice. Much is known about factors that influence assessments, and different solutions claim to improve the validity and reliability of assessments of students' clinical competency. However, little is known about *how* teachers go about improving their assessment practices. This study aims to contribute empirical findings about how teachers' assessment practice may change when shared criteria for assessing students' clinical competency are developed and implemented.

Methods: Using a narrative-in-action research approach grounded in narrative theory about human sense-making, one group including nine health professions teachers was studied over a period of 1 year. Drawing upon data from observations, interviews, formal documents and written reflections from these teachers, we performed a narrative analysis to reveal how these teachers made sense of experiences associated with the development and implementation of joint grading criteria for assessing students' clinical performances.

Results: The findings present a narrative showing how a shared assessment practice took years to develop and was based on the teachers changed approach to scrutiny. The teachers became highly motivated to use grading criteria to ensure fairness in assessments, but more importantly, to fulfil their moral obligation towards patients. The narrative also demonstrates how these teachers reasoned about dilemmas that arose when they applied standardised assessment criteria.

Discussion: The narrative analysis shows clearly how teachers' development and application of assessment standards are embedded in local practices. Our findings highlight the importance of teachers' joint discussions on how to interpret criteria applied in formative and summative assessments of students' performances. In particular, teachers' different approaches to assessing 'pieces of skills' versus making holistic judgements on students' performances, regardless of whether the grading criteria are clear and well-articulated on paper, should be acknowledged. Understanding the journey that these teachers made gives new perspectives as to how faculty can be supported when assessments of professionalism and clinical competency are developed.

1 | INTRODUCTION

The assessment of student learning in health professions' education is a central, yet challenging task.[1,2] One challenge involves achieving transparency through the application of pre-defined standards, while also acknowledging teachers' professional judgements. For the past decades, outcome-based and competency-based curriculum reforms have pushed for enhanced transparency and public accountability.[3–5] One way to achieve such transparency has been to make the assessment standards explicit.[3,5] However, the development of meaningful criteria capturing competency, assessment validity and reliability are debated.[5–9] Explicit grading criteria are known to increase the transparency of what students are expected to learn but may endanger the assessment of integrated competency in favour of 'pieces of' competencies.[10–12] While much is known concerning the outcomes of different assessment methods,[13] and the reasons for variation in assessor ratings including proficiency in making judgements and providing feedback,[14,15] little is known about *how* clinically oriented teachers develop assessment practices and make sense of assessment standards. Gordon and Cleland[16] recently called for non-linear approaches to understand change practices in context. This narrative study uses such a non-linear approach to unravel the complexity of change practices by contributing empirical-based findings concerning how health professions teachers go about their assessment practice, over time, and in relation to policies emphasising clear assessment criteria. The aim of the present study is to understand *how teachers' assessment practice may change when shared criteria for assessing students' clinical competency are developed and implemented*.

1.1 | Teachers' approaches to assessment and standards

There are a number of explanations as to why assessors' ratings differ, for example that student performance is judged based on social categorisations of individual charachteristics.[14] Kogan et al.[17] report several factors leading to variation between faculty members' assessments of clinical skills. They emphasise the influence of contextual factors in the assessment situation, such as the educational setting, the unique clinical encounter and the institutional culture.[17] Teachers' epistemological views are also known to influence their assessment practices.[4,18,19] Such fundamental assumptions 'come to life' and are an integrated part of the social and cultural context in which health professionals practice.[20] Enacted views may explain the variation between assessors' judgements, and also, research suggest that teachers regard the function of assessment in different ways.[15,18,21] de Jonge et al.[21] identified key themes in the literature regarding different perspectives of work-based performance assessments: (1) assessment *for* learning versus assessment *of* learning, (2) holistic versus analytical conceptualisations of competence and (3) psychometric versus social-constructivist approaches. Hodges[22] details how the psychometric discourse has not only dominated the medical education research regarding assessments but also how it has been a way of thinking and practising among educators, for example, by providing feedback using numbers. The use of numbers rather than words reflects philosophical assumptions and may save time, but researchers urge caution and suggest a combination of rating and feedback methods could be valuable and serve different purposes.[23] Using a similar rationale, advocates of programmatic assessment suggest a variety of formative and summative assessment methods over longer periods of time that capture students' capabilities in various ways.[1,24]

While much research concern the identification of explanatory factors[7,14,17] and successful methods for assessments,[13,25] little is devoted to how teachers' views may change over time. In this paper, we address teachers' development of practices and apply O'Donovan et al.'s[19] framework, which outlines teachers' different approaches to developing students' understanding of assessment standards, the *laisse-faire* approach, the *explicit* approach, *the social-constructivist* approach *and the community of practice* approach.[19] Practising the *laisse-faire* approach to assessment means students gradually come to 'know' how standards are set and how quality is assessed. Teachers with a laissez-faire approach judge performances according to tacit standards that are informally communicated in serendipitous ways. The *explicit* approach is characterised by assessment criteria that articulate standards explicitly but passively. This approach has been criticised for making teachers rely too much on so-called explicit criteria. The *social-constructivist* approach acknowledges joint participation with respect to evaluative practice. Students are actively engaged and, through various activities, become familiar with assessment criteria to create an understanding of what they mean in practice.[6] The fourth approach builds on Wenger's[26] theory of *community of practice*. It acknowledges the importance of teachers' and students' mutual engagement in the assessment practice, whereby explicit standards and tacit knowledge within the learning community are discussed and shared to form mutual understanding.

2 | METHODS

2.1 | A narrative research approach

This study is based on narrative theory about how humans bring meaning to their experiences by structuring them as narratives.[27–29] Narrative sense-making means that people connect past, present and future events into narratives to explain actions and experiences.[30,31] Bruner[27] argues that there are two complementary ways to make sense of the world: the logico-scientific and the narrative. Science is underpinned by the first, in which logic is used to find general causes, and through empirical explorations test verifiable truths. Narrative sense-making, however, deals with how humans explain the actions of themselves and others by making connections between different events in their everyday lives.[27] When these storied connections evolve, they may contain contradictions and multiple meanings.[28] The rationale of this study is that development of practice is filled with contradictions and that a better understanding of how teachers make sense of experiences related to change is essential to improve educational practice. The strength of narratives is their lifelikeness, which illuminates the messy and contradictory aspects of human life.[27] When teachers engage in the diversity of academic and clinical work, their intentions are not static or the result of 'a competence development'. Rather, intentions change continuously as different ways of practising are tested and reflected upon.

2.2 | Context of study and sampling

This study was conducted in conjunction with national reforms stressing transparency such as clear grading criteria within higher education in Sweden. Universities in Sweden have the autonomy to decide how grading criteria are applied, and therefore, there is variation between universities, courses and even within departments. Study programmes in Sweden are organised around a course-based system, where students' performances are assessed and graded after each course, which generally lasts for 5 or 10 weeks. In Sweden, course leaders are mandated to develop the course syllabus, decide assessments and grading criteria and usually have the formal role of examiner, but other teachers can provide input on student performance. In medical and health professions education, each syllabus should specify learning outcomes that are linked to the curriculum and thus to the intended graduate competency.[4] The level of detail and the way that learning outcomes are expressed varies between courses but, in general, grading criteria specify the requirements on student performance.

In accordance with our in-depth narrative research approach, one educational setting was chosen using theory-based sampling.[17] We recruited nine teachers who had implemented major curriculum changes where they translated policy based on a broad interpretation in line with a student-centred view of learning, similar to what is described by the theoretical construct *ideological approach* to curriculum reform.[32] All nine teachers, men and women, worked at a hospital site and were responsible for the planning, delivery and evaluation (including formative and summative assessments) of the main courses in one health professions education offered by a Swedish university. In the last 10 years, responsibility for the majority of courses in the study programme had rotated between these teachers, who each had been course leaders and examiners for several courses.

This study was conducted in conjunction to an intervention initiated by the teachers to enhance their assessment practice. The teachers recruited three students from different levels of study and videotaped them as they each examined different patients. The examination included history-taking, physical examination, handling technical devices, interpretation of findings before diagnosis and recommendations for treatment. The examination was expected to take approximately 1 hour and was performed on authentic patients in a clinical training setting that resembled the students' final clinical skills exam. The teachers then met on three occasions at 5-month intervals. During the first two meetings, all teachers watched the videos and carried out the assessments individually before jointly discussing their assessment outcomes and interpretations of the criteria. Both meetings resulted in refinements of the criteria, such as multi-level rankings and clarifications on professional behaviour. Five months after their second criteria discussion, the teachers met again to finalise the criteria template. In between meetings, they applied the revised criteria in their practice.

2.3 | Data and analysis

The data were generated through a combination of methods throughout the teachers' 1-year intervention to enhance their assessment practice. Tape-recorded and transcribed observations were made during four meetings of the nine teachers and during informal talks between those meetings. Their individual written reflections following the meetings were also collected. The field notes generated by the first author included facial expressions, body language, the physical room and artefacts and the atmosphere.[33,34] The notes were all written out either the same day or the day after the observation. At the end of the year, three of the teachers who had worked the longest at the department (>10 years), and the teacher responsible for coordinating the revision of the grading criteria, were chosen for a group interview. This exploratory interview (2.5 hours) provided an opportunity for the teachers to recall past experiences and evolve the meaning of these, thereby rich data was generated.[33] Two additional interviews were held with the Programme Director, who was also part of the teacher group (e.g. course lead and examiner).

Three of the authors jointly conducted a narrative analysis centred on *significant events*,[28,29,34] i.e. events the teachers perceived as significant[28] for their joint development of assessments, either by creating opportunities or pressure to change. These overlapping events, see Table 1, were based on stories shared by the teachers about situations that went as far back as 10 years. During analysis, all data, transcribed as text materials, were pooled together, enabling narrative analysis of *how* the teachers' 'prevailing discourses' were

TABLE 1 Significant events

Significant events

- University reform with increased requirements of transparency, followed by curriculum revision.
- Students' performances deteriorated, and they complained about too little support and wanted to drop out.
- Competence development in pedagogy resulted in the implementation of peer-assessments and mini-CEX.
- Reduction in the number of assessors grading students' clinical skills from two to one.
- Introduced regular teacher-meetings to discuss educational matters.
- Students complained about unfairness in assessments.

expressed in their everyday enactment (thinking and acting) and how these discourses evolved over time.[34] The analytical process conducted is known as emplotment, which means that the researchers linked human action, meaning, motives, events and consequences in the 'same way' the teachers did, in order to make sense of their experiences.[29,34] Plots are ordered around a beginning, middle and end, which makes the findings from a narrative analysis more than a brief outline of human reasoning and differs from the presentation of a thematic analysis of narratives.[30,35] Plots may be structured around how different events played out in a physical sense, however, as we attended to human experience of change, the emplotment centred on happenings the teachers expressed as meaningful. In accordance with our research approach, human meaning-making was regarded as a recreation of time and interconnected events that most often differ from physical chronology.[27,28] The unfolding narrative (the emplotment) was rewritten several times and discussed among the authors. Eventually, a coherent narrative was structured, depicting the teachers' meaning-making (shown through action and experience) of how their assessment practices had changed and how they made sense of the criteria to assess students' clinical competency.

2.4 | Methodological reflections and limitations

Narratives are embedded in social contexts and therefore unique and not meant to be generalised. They depict the richness and complexity of a phenomenon, and what unique narratives illustrate can be transferred to explain and understand happenings in other contexts, and for that purpose, contextual descriptions are included here.[33] The narrative reported here illustrates change processes including motives and events that were meaningful to the teachers. From a natural science perspective, humans' recollection of events may be biased; for example, narrative sense-making may not reflect a precise chronological presentation of events. Narrative-in action analysis thus illuminates enacted stories and how humans make sense of events *from their perspective*. As we adopted a socio-cultural perspective,[26,36] and attended to the group level as the unit of analysis, the teachers' individual differences in sense-making are not addressed here. The analysis included data generated from nine teachers who conducted an intervention and had the main responsibility for one study programme, although other health practitioners and university faculty who taught and assessed their students may not have shared their perspectives. Furthermore, the participant teachers had previously attended faculty development and assumed to be pedagogically informed, although not all of them had training specifically regarding assessment.

2.5 | Ethical considerations

In accordance with the ethical approval for this study, all participants formally consented to take part after being informed orally and in writing. The teachers are here given pseudonyms, and to further ensure confidentiality, no details concerning professional activities are disclosed.

3 | RESULTS

3.1 | Safeguarding fairness in assessments

The findings present a narrative where the teachers became motivated by moral intentions to calibrate their use of grading criteria for assessing students' clinical skills. In this section, the prologue first explains, based on our analysis, the teachers' intentions of reworking assessment criteria and the curriculum. Then, the narrative outlines (a) how the teachers developed an assessment-oriented culture in which criteria were embedded, (b) the ways in which the teachers made sense of those and (c) how the teachers' development of grading criteria took different turns in connection with their values of fairness in assessments. The epilogue then shows how the teachers summarised their intervention. Included in the narrative, the teachers' ongoing dialogues and accounts situate and depict how their endeavours became manifest. These accounts include short stories that were shared among the teachers in and between meetings or in interview situations.

3.2 | Prologue: Motives to initiate change

A group of health profession teachers working together at a hospital site had conducted a significant curriculum reform towards outcome-based education. As part of this change process, they increased the emphasis on training of clinical skills, re-defined their teaching roles, adopted a facilitating role and reduced their time as information providers. The increase of assessments concerning clinical skills forced them to economise resources from two examiners to one. This felt a bit unreliable, so they developed joint criteria useful for both formative and summative assessments. However, the students increasingly complained that they were being assessed unfairly and that some teachers were making harsher judgements. The teachers at first rejected these complaints, but during post-assessment meetings, they realised that they had different understandings of the criteria and different ways of judging student performance. This led them to conduct an intervention with the aim of harmonising the application of the criteria and assessment of students' clinical skills.

3.3 | How the teachers opened up to scrutiny

3.3.1 | The narrative starts in present time and outlines how a decade of changes made joint development of grading criteria possible

The ideas behind improvements to the assessment criteria arose from the teachers' practice as it evolved over the last 10 years. By opening up to scrutiny, they paved the way for a shared understanding that applying clear assessment criteria was a moral obligation. The following story depicts how the use of criteria had become an integral part of their approach to support student learning through assessments.

> On his way to a meeting, John was stopped by a student who stuck her head out of one of the training rooms and beckoned him. She said, ‘Hey John, would you like to watch while I examine Anna?’ John, who still had some time before his meeting, was happy and relieved by her invitation. The student was known to be shy and reluctant to participate in continuous assessment. Sometimes, it was hard to tell if she was insecure about the patient examination procedures or if she was just uncomfortable having her performances scrutinised. Pleased that she had finally opened up, he smiled at her and said: ‘I'd be happy to!’

As illustrated by the above story, the everyday spirit at the department was open and friendly. Through the teachers' persistent work over several years, and illuminated by the narrative analysis, they had developed an assessment-oriented culture, which was shown in how peer learning and evaluations were continuously applied. In several ways, the teachers revealed how dialogue and peer assessments had developed into a habitual practice among the students and among themselves. They spoke openly about how they felt comfortable with continuous evaluations of their teaching and of jointly making educational improvements.

Our narrative analysis shows that the teachers' efforts to jointly develop grading criteria and make assessments fair were the result of a change process in which they gradually opened up to scrutiny. The implementation of assessment criteria was preceded by years of trial and error in applying different teaching methods and learning activities. Peer assessments started out, about 5 years ago, as one way to support students taking greater responsibility for their learning. After a few years of applying peer learning, it became commonplace for students to invite others to provide feedback while practising clinical work. The teachers believed, they ‘had gained a lot’ by opening the door to continuous evaluation as it created space for creativity and new developments. However, it was the teachers' belief that the application of explicit assessment standards represented the moral good that drove their ambitions further.

3.4 | How the teachers enacted the ‘moral good’

3.4.1 | This second part of the narrative illustrates why grading criteria became meaningful to the teachers and how past enactments reinforced current initiatives to assess fairly

As the teachers had taken several initiatives to improve their assessment practice and provide a high quality education, caring for patients and students was their key motivation. Ultimately, the assessment criteria were means for teachers to fulfil their obligations towards patients by ensuring graduates had the necessary competency. For example, their idea of time-limited assessments had little to do with effectiveness in professional work, but rather concern of patients' discomfort during physical examination, which students needed to minimise and thereby preventing patients from being afraid to seek help. The teachers' efforts to enact the ‘moral good’ were reinforced and justified by stories about past experiences of student assessment, such as Hanna's recollection of high-stake final exams.

> During the course of your life, you are never evaluated on your practical skills. The only time is when you take your driving test. No wonder the students lacked experience of being assessed on their performance! It was really unethical, when you think about it. They pursued their studies over several years, and right before they graduated, they were graded on their clinical performances. Some of them had already got jobs. And then, bang, they failed their final exam! They were so nervous their faces turned green, and they were ready to faint.

Such stories were shared repeatedly among the teachers and deepened the perception of how assessment practices in the past were inadequate, compared with the present system. In this way, past events confirmed how their reformation of the curriculum including adding continuous clinical skills training was morally justifiable. The teachers believed these changes had led to students feeling better prepared and performing better. They had also come to understand that being assessed on practical skills was a new and highly stressful situation for the students and something that should be regarded as an ability in its own right that required training and evaluation. However, with the curriculum changes and the use of continuous peer learning and feedback, the graded assessments had become less dramatic.

Applying criteria to the assessment of clinical skills became meaningful because it enforced the teachers' values of being fair and ensuring patient safety. However, enacting the narrative about the moral good meant that the values of fairness in assessment, patient accountability and facilitation of learning sometimes collided. It also presented the dilemma of deciding what was fair and what was not.

3.5 | How the teachers upheld fairness with an unbiased assessment

3.5.1 | This third part of the narrative shows how the joint development of grading criteria took different turns associated with the teachers' value of assessing students fairly

During the discussions of how to harmonise assessments, the teachers' efforts to apply criteria in ways that promoted certain student behaviours became clear. One way to ensure that the students met the minimum requirements of patient safety and good practice was to define absolute requirements of what they should and should not do when examining patients, such as washing their hands and disinfecting instruments. Performances like that were assessed pass/fail, regardless of the varied quality of how students carried out these tasks, consequently, either the student performed these tasks or the student would fail the entire exam. On the one hand, these 'either-or' performances were seen as easy to assess; on the other hand, there was concern when students performed such activities partially, as in the case of a student who cleaned a few of the instruments, but not all of them. One suggestion was that the teachers could take into account that all of the instruments actually used to examine the patient had in fact been disinfected. According to the criteria, however, all hygiene aspects were stipulated as non-negotiable, which made it reasonable to fail students who neglected to disinfect all instruments regardless of whether they had been used in the examination. Some teachers argued in favour of this type of assessment, contending it was easier to conduct, non-negotiable and therefore fair. Behind their argument was a concern for patients and that a student who neglected to disinfect all instruments could not be fully trusted to treat patients. This non-negotiable 'either-or' reasoning was also adopted when they applied criteria to assess performances of a different nature, as illustrated by the dialogue below.

John: If everyone agrees that the student never made a summary of the history-taking, then how come we all graded the student as pass when the criteria clearly says that this *should* be done?

Tina: Hmm, very good question!

Edward: But what she did, the things she performed, she did really well. She just never really got it completely.

Hanna: I think we need to split this criterion into two parts, otherwise it will be hard to give feedback. The first part should be the technique used during the procedure, so that, in this case we can give some credit for all that she did. I feel that would be fair. And then, the summary of the history-taking can be a separate criterion.

Beatrice: If we do as you suggest, should the summary then still count as five points and be a criterion we use to bring them down?

Jenny: You mean if they haven't made a summary of the history-taking they fail?

Hanna: Well, yes, as long as it says *should*, here in the template.

Edward: In the eyes of the students, it will be clear that you fail your exam if you don't do this!

Hanna: Yes, and they do what the template says!

John: Ok, so if they don't summarise the history-taking, they fail the whole exam?!

Hanna: Well yes, if they fail to summarise, we will never know if they understood why the patient came in the first place. Did you handle the problem correctly? Well, there is no way of knowing if you never identified the problem in the first place.

During the teachers' work to develop the criteria, they reflected upon how assessments had been performed in the past. Recurrent comments during the discussions highlighted how the teacher group had changed from making subjective assessments of student performance in the past, to being as objective as possible. Together, they laughed about the lack of explicit criteria back in the 'old days' when a former professor once said about a student's examination, 'She is so cute, she can pass'. However, being entirely objective was considered difficult when using the standardised criteria to assess student–patient encounters of a different nature and during the debates the teachers repeatedly reminded each other 'but then it becomes subjective again'. Subjectivity was mainly associated with the assessment of professional behaviour and communication skills, and therefore, clarifying those criteria created a need to define performance dimensions that could not be misinterpreted. Trying to, in various ways, defining dimensions of professional behaviour, the teachers reasoned about the differences between behaviour, overall communications, the summary of history-taking, giving information about the diagnosis and treatment and the use of jargon. They teetered between two different rationales on the assessment of professional behaviour: assessing overall communication and professional behaviour or dividing the communication into separate pieces and connecting it to each part of the patient examination. They agreed that communication skills were somewhat different from professional behaviour and could perhaps be assessed separately.

In an attempt to be fair and consistent in assessments, multiple interpretations of criteria and student performances were scrutinised by the teachers. However, they believed that standardisation was not fully compatible with the reality of clinical work, for example, when students examined patients that were considered particularly troublesome.

Beatrice: How do we explain to students who failed because the examination took too long, when, at the same time they have a friend who passed who also exceeded the time limit? Should we perhaps add the ten percent time margin that we use for the mini-CEX?

George: No! You cannot let yourself be steered by the template that hard!

Jenny: Agree. We're not robots!

John: I agree. If that is the case, I mean if you see that the patient is being particularly difficult, you just have to commit a criminal act and deviate from the template.

The dialogue above shows how the teachers tried to achieve fairness by following the criteria template and recognised reality as multifaceted and that making professional judgements required considering the complexity of the situation.

3.6 | Epilogue—No perfect assessment criteria

At the end of their intervention, the teachers reflected upon how hard it was to create equal situations for the students and came to the conclusion that there was no such thing as perfect assessment criteria. Reflecting on their intervention, they felt that even though they valued single criteria differently, the overall assessment of each student was more equal than they had anticipated. This made the teachers conclude that, even if assessments would never be completely harmonised, their joint discussions had led to a negotiable consensus.

4 | DISCUSSION

The findings show how the teachers' development of common grading criteria was made possible by their openness to peer scrutiny and that these changes were driven by their values of fairness and accountability. Criteria had been integrated in the curriculum, useful for formative and summative assessments. In that way, and from the teachers' perspective, criteria safeguarded fair assessments and that future patients would receive the best possible treatment. However, a number of dilemmas emerged, such as the assessment of integrated competencies versus the assessment of separate ‘pieces of skills’.

The establishment of an assessment-oriented culture seemed to be prerequisite for how the teachers were committed to harmonise their application of grading criteria. This change in assessment practice, from tacitly conveyed expectations to shared understandings of criteria, can be understood vis-à-vis O'Donovan et al.'s[19] model of teachers' approaches to sharing standards. By referring to how tacit standards, as in the *laisse-faire approach* prevailed in the past, the teachers justified their choice to apply clear criteria. Grading criteria were then implemented, in response to tacit standards, but the teachers realised that articulating criteria were not enough. Individual students' understanding of assessment standards may differ,[19] and the teachers in the present study came to realise that, to fulfil their intentions concerning assessments, all students needed to make sense of the criteria in the same way. Thus, formative assessments in parallel with students' peer reviews were integrated throughout the curriculum, which enabled the enactment of a *social-constructivist approach*.[19] Interestingly, it appears that involving students in the dialogue about applying grading criteria created a need for further clarifications. It seems plausible to conclude that, when students have full access to the standards by which they are judged, there is an opportunity to discuss their performances in light of these standards. Consequently, teachers may need to reflect on ways to interpret criteria, the range of acceptable student performances and how to justify their judgements.

We agree with Kogan et al.[17] that shared standards articulated via, for example, a criterion-referenced framework can mediate feedback. While explicit criteria can facilitate learning, they say nothing about the quality of those standards and, therefore, do not safeguard that, for example, teachers' judgements are valid.[1] As others point out,[1] validity and reliability are not immanent traits of tests and will not be achieved simply by applying an assessment instrument. Joint negotiations within teacher communities—similar to the discussions held by the teachers in this study—will likely harmonise the understanding of both criteria and judgements of student performance. The teachers in this study did not invite their students to be co-participants in formulating assessment criteria, as in the *community of practice* approach.[19,26] However, their efforts can be understood, in part, as a shared community of practice around assessment matters. Through a development process, the teachers opened themselves up to each other's ideas and critiques, which enabled negotiations about a shared meaning of the grading criteria, competencies/competency and assessment. The problem of formative assessments being taken less seriously by students and teachers[24] seemed to be avoided by the teachers' efforts to align low and high stake assessments and to embed continuous teacher and peer feedback as part of an assessment-oriented culture. However, this change process took time and included shifts in assessment rationales.

The teachers' ambitions show how dilemmas in assessment manifested, such as inter-rater reliability and standardisation, and acknowledging contextual factors arising in patient encounter. The idea of being steadfast to the criteria collided with the notion of sometimes having to deviate from the template, and they agreed that particularly complex patient cases should be taken into account before grading. In a similar way, Kogan et al.[17] report that faculty members' are influenced by the complexity of clinical encounters when making performance ratings, arguing that faculty needs to be trained in assessment to modify such rating errors.[17] While we agree that faculty development is beneficial, the challenge to decide what counts as valid in a given context still remains.[25,37] Moreover, and shown by this study, as assessment practices change, and teachers translate their ‘new knowledge’ into practice, they may face new dilemmas. The current findings imply that faculty development needs to address how assessors make sense of criteria and to involve clinical teachers in joint discussions on the range of acceptable student performances, which seem to harmonise teachers' ratings.

In order to achieve equal and unbiased assessments, the teachers in this study wondered whether separate pieces of student performance should be stated in the grading criteria. Consequently, they tried to operationalise holistic criteria (competency) by, for example, splitting professional behaviour into subcategories that could be judged binarily. Thus, although they adopted a social-constructivist approach to students' understanding of standards,[19] the teachers enacted a different rationale to develop the grading criteria. The challenges of constructing reliable measurements of clinical competency have been acknowledged,[38] though, relying solely on the use of binary checklists based on psychometric rationales has also been questioned.[12,39] The critique related to criterion-referenced assessments claims teachers risk judging pieces of performance rather than integrated competencies useful for professional practice.[8] Global ratings of professionalism were seen by the teachers in this study as being subjective and incompatible with their intentions of protecting patients and being fair to students. Our interpretation is that the teachers abstained from assessments that were uninformed or based on tacit aspects, similar to the laissez-faire approach.[19] Holistic judgements need not be confused with making biased or invalid judgements,[22] and objectivity is not equal to reliability obtained through complete standardisation.[1,39] The question the teachers in this study raised was whether 'something was lost' in the details. It has previously been established that teachers need to synthesize student achievements before grading; 'When we see the whole, we see its parts differently than when we see them in isolation'[4(p227)]. While certain performances in professional health practices need to be non-negotiable due to patient safety, criteria of professional behaviour, for example, may be less valid if binary judgements are made on separate dimensions. The teachers' reasoning in this study made visible how tempting it may be to use the same rationale when formulating criteria for performances of very different types. Consequently, the assessment of integrated competencies may be lacking. This implies that teachers should reflect on how standards and ratings of clinical competency need to acknowledge the simultaneous use of different rationales, thus a combination of binary and holistic judgements.

Educational change and teachers' development of assessment practice are often regarded as slow and resistant.[18,21] This study shows how teachers' motivation to develop grading criteria was derived from their own practice working with students and their concern for patients. Thus, their willingness to do good for society and for the students was the incentive for creating a shared assessment practice, which had little to do with pressure from the university or governmental reform stressing transparency.

5 | CONCLUSION

This study contributes a rich description of how teachers' assessment practices may change when shared assessment criteria are developed and implemented. The change process illustrated in this paper neither stipulates neither a linear model nor an ideal development process, yet a number of implications may be drawn from this study. The findings imply that teachers need to regularly re-evaluate grounds for their judgements through joint discussions of criteria and the range of acceptable student performances. Such discussions seem to not only harmonise the understanding of criteria application but also unravel the shifting rationales on assessment and competency within teacher communities. This study demonstrates the adaptation of peer learning and social-constructivist approaches may take time and create new choices and dilemmas in assessment. Whereas some researchers argue for increased rigour in performance tests, others call for holistic, constructivist and professional approaches or suggest triangulation of assessments over longer periods of time. Regardless of what kind of assessments is applied, conversations that take charge of teachers' professional judgements are necessary. This study shows that teachers' views on assessment are not fixed, and they should reflect on how assessment standards and their judgements must acknowledge the simultaneous use of different rationales. Therefore, with reference to how individual student performances involving patients with various needs should to be judged differently, a combination of binary and holistic judgements needs to be applied. We welcome more research on assessment practices beyond individual teachers' views and on how teachers make sense of applying different methods and standards.

ACKNOWLEDGEMENTS

We wish to thank the teachers who participated and shared their practice and the colleagues who provided valuable comments on the manuscript. This work was supported by the Stockholm County Council, grant no. SLL 20110191.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

AUTHOR CONTRIBUTIONS

All authors meet 1–4 criteria and have given their approval to submit this manuscript; more specifically, Dr Linda Barman has been the main researcher during design, data collection and analysis including interpretation, as well as writing of this paper. Ass. Prof. Cormac McGrath has contributed in the analysis (emplotment) and writing of the paper. Prof. Staffan Josephsson has as a method expert been essential and consulted during the design, and he contributed in the analysis including the interpretation, as well as critically revising of the manuscript. Ass. Prof. Charlotte Silén has substantially contributed to the study design and critically revising of manuscript. Prof. Klara Bolander-Laksov has substantially contributed to the study design and the analysis including interpretation, as well as critically revising the manuscript.

ETHICS STATEMENT

Approval was based on supplementary application from the Regional Ethics committee in Stockholm, Sweden. The supplementary application was deemed exempt from full review (no 418-32).

ORCID

Linda Barman https://orcid.org/0000-0001-7148-3271
Cormac McGrath https://orcid.org/0000-0002-8215-3646
Klara Bolander Laksov https://orcid.org/0000-0002-3345-3810

REFERENCES

1. van der Vleuten C, Schuwirth L. Assessing professional competence: from methods to programmes. *Med Educ*. 2005;39(3):309-317. doi:10.1111/j.1365-2929.2005.02094.x
2. Epstein RM. Assessment in Medical Education. *N Engl J Med*. 2007;356(4):387-396. doi:10.1056/NEJMra054784
3. Morcke A, Dornan T, Eika B. Outcome (competency) based education: an exploration of its origins, theoretical basis, and empirical evidence. *Adv Health Sci Educ*. 2012/09/01;2012(4):1-13. doi:10.1007/s10459-012-9405-9

4. Barman L, Silén C, Bolander Laksov K. Outcome based education enacted: teachers' tensions in balancing between student learning and bureaucracy. *Adv Health Sci Educ*. 2014/01/24;2014(5):1-15. doi: 10.1007/s10459-013-9491-3
5. Hodges BD. A tea-steeping or i-Doc model for medical education? *Acad Med*. Sep 2010;85(9 Suppl):S34-S44. doi:10.1097/ACM.0b013e3181f12f32
6. Carraccio CL, Englander R. From flexner to competencies: reflections on a decade and the journey ahead. *Acad Med*. 2013;88(8):1067-1073. doi:10.1097/ACM.0b013e318299396f
7. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box' differently: assessor cognition from three research perspectives. *Med Educ*. 2014;48(11):1055-1068. doi:10.1111/medu.12546
8. Talbot M. Monkey see, monkey do: a critique of the competency model in graduate medical education. *Med Educ*. 2004;38(6):587-592. doi:10.1046/j.1365-2923.2004.01794.x
9. Govaerts M. Workplace-based assessment and assessment for learning: threats to validity. *J Grad Med Educ*. 2015;7(2):265-267. doi:10.4300/jgme-d-15-00101.1
10. van der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. *Adv Health Sci Educ*. 1996;1(1):41-67. doi:10.1007/bf00596229
11. Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA*. 2002;287(2):226-235. doi:10.1001/jama.287.2.226
12. Hodges B. Medical education and the maintenance of incompetence. *Med Teach*. 2006;28(8):690-696. doi:10.1080/01421590601102964
13. Weller JM, Misur M, Nicolson S, et al. Can I leave the theatre? A key to more reliable workplace-based assessment. *BJA: British Journal of Anaesthesia*. 2014;112(6):1083-1091. doi:10.1093/bja/aeu052
14. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: rethinking the etiology of rater errors. *Acad Med*. 2011;86(10):S1-S7. doi:10.1097/ACM.0b013e31822a6cf8
15. Berendonk C, Stalmeijer RE, Schuwirth LWT. Expertise in performance assessment: assessors' perspectives. *Adv Health Sci Educ*. 2013;18(4):559-571. doi:10.1007/s10459-012-9392-x
16. Gordon L, Cleland JA. Change is never easy: how management theories can help operationalise change in medical education. *Med Educ*. 2021;55(1):55-64. doi:10.1111/medu.14297
17. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model. *Med Educ*. 2011;45(10):1048-1060. doi:10.1111/j.1365-2923.2011.04025.x
18. Boud D, Dawson P, Bearman M, Bennett S, Joughin G, Molloy E. Reframing assessment research: through a practice perspective. *Stud High Educ*. 2018;43(7):1107-1118. doi:10.1080/03075079.2016.1202913
19. O'Donovan B, Price M, Rust C. Developing student understanding of assessment standards: a nested hierarchy of approaches. *Teach High Educ*. 2008;13(2):205-217. doi:10.1080/13562510801923344
20. Whitehead CR, Austin Z, Hodges BD. Continuing the competency debate: reflections on definitions and discourses. *Adv Health Sci Educ*. 2013;18(1):123-127. doi:10.1007/s10459-012-9407-7
21. de Jonge LPJWM, Timmerman AA, Govaerts MJB, et al. Stakeholder perspectives on workplace-based performance assessment: towards a better understanding of assessor behaviour. 2017;22(5):1213-Advances in Health Sciences Education, 1243. doi:10.1007/s10459-017-9760-7
22. Hodges B. Assessment in the post-psychometric era: Learning to love the subjective and collective. *Med Teach*. 2013;35(7):564-568. doi:10.3109/0142159X.2013.789134
23. Ginsburg S, Watling CJ, Schumacher DJ, Gingerich A, Hatala R. Numbers encapsulate, words elaborate: toward the best use of comments for assessment and feedback on entrustment ratings. *Acad Med*. 2021;96(7S):S81-S86. doi:10.1097/ACM.0000000000004089
24. Schut S, Heeneman S, Bierer B, Driessen E, van Tartwijk J, van der Vleuten C. Between trust and control: teachers' assessment conceptualisations within programmatic assessment. *Med Educ*. 2020;54(6):528-537. doi:10.1111/medu.14075
25. van der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205-214. doi:10.3109/0142159X.2012.652239
26. Wenger E. Communities of Practice. In: *Learning, Meaning and Identity*. Cambridge University Press; 2008.
27. Bruner J. *Actual minds, possible worlds*. Harvard University Press; 1986. doi:10.4159/9780674029019
28. Mattingly C. *Healing Dramas and Clinical Plots: The Narrative Structure of Experience*. Cambridge University Press; 1998. doi:10.1017/CBO9781139167017
29. Mattingly C. In search of the good: narrative reasoning in clinical practice. *Medical Antropology Quaterly*. 1998;12(3):273-297. doi:10.1525/maq.1998.12.3.273
30. Bleakley A. Stories as data, data as stories: making sense of narrative inquiry in clinical education*. *Med Educ*. 2005;39(5):534-540. doi:10.1111/j.1365-2929.2005.02126.x
31. Josephsson S, Asaba E, Jonsson H, Alsaker S. Creativity and order in communication: Implications from philosophy to narrative research concerning human occupation. *Scand J Occup Ther*. Jun 2006;13(2):86-93. doi:10.1080/11038120600691116
32. Barman L, Bolander Laksov K, Silén C. Policy enacted—teachers' approaches to an outcome-based framework for course design. *Teach High Educ*. 2014;19(7):735-746. doi:10.1080/13562517.2014.934346
33. Denzin NK, Lincoln YS. *Collecting and Interpreting Qualitative Materials*. 2nded. SAGE Publications; 2003.
34. Josephsson S, Alsaker S. Narrative Methodology: a tool to access unfolding and situated meaning in occupation. In: Nayar S, Stanley M, eds. *Qualitative Research Methodologies for Occupational Science and Therapy*. Routledge; 2015:70-83.
35. McCance TV, McKenna HP, Boore JRP. Exploring caring using narrative methodology: an analysis of the approach. *J Adv Nurs*. 2001;33(3):350-356. doi:10.1046/j.1365-2648.2001.01671.x
36. Bleakley A. Socio-cultural learning theories. In: Bleakley A, Bligh J, Browne J, eds. *Medical Education for the Future Identity, Power and Location*. Springer Science+Business Media; 2011:43-60 Advances in Medical Education.
37. 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):239-260. doi:10.1007/s10459-006-9043-1
38. Shumway JM, Harden JR. AMEE Guide No 25: the assessment of learning outcomes for the competent and reflective physician. *Med Teach*. 2003;25(6):569-584. doi:10.1080/0142159032000151907
39. Schuwirth LWT, van der Vleuten CPM. A plea for new psychometric models in educational assessment. *Med Educ*. 2006;40(4):296-300. doi:10.1111/j.1365-2929.2006.02405.x

How to cite this article: Barman L, McGrath C, Josephsson S, Silén C, Bolander Laksov K. Safeguarding fairness in assessments—How teachers develop joint practices. *Med Educ*. 2022;56(6):651-659. doi:10.1111/medu.14789