WMU Journal of Maritime Affairs (2020) 19:493–508
https://doi.org/10.1007/s13437-020-00210-2

IAMU SECTION ARTICLE

Pedagogical dilemmas in dynamic assessment situations: perspectives on video data from simulator-based competence tests

Charlott Sellberg [1]

Received: 13 April 2020 / Accepted: 22 June 2020 / Published online: 30 June 2020

Abstract

This study investigates navigation instructors' explanations of dynamic assessment practices during simulator-based competency tests, adopting a video-stimulated recall method. Episodes of authentic video materials from simulator-based competency tests are selected and subjected to interaction analysis. In the next step, the episodes are used for conducting stimulated recall interviews with navigation instructors ($n = 11$) in two focus groups. The results reveal the dynamic nature of assessing competence as well as how instructors participating in focus groups identified and critically discussed a variety of pedagogical dilemmas related to these dynamics. These are related to aspects that relate to what constitutes a valid exam question, how to assess students' responses, and consistency and fairness of competence tests. In particular, the results indicate the complexity of conducting valid and reliable assessments of knowledge-in-action in situ as well as how thoughtful scenario designs could reduce inconsistencies and unequal treatment of students. The results also highlight how a repeated and collaborative viewing of videos was helpful for instructors to identify potential problems in the dynamic assessment situations they viewed. The implications of the results highlight the need for conducting high-stake assessments for maritime certificates based only on observable behavior and video records of competence tests rather than during ongoing simulator tests. Lastly, the need for continuous and structured pedagogical development of instructors is identified in order to support their complex work of training and assessing competence.

Keywords Maritime education and training (MET) · Simulator-based competency tests · Dynamic assessment · Interaction analysis · Video-stimulated recall · Pedagogical development

✉ Charlott Sellberg
charlott.sellberg@ait.gu.se

[1] Department of Applied Information Technology, University of Gothenburg, Forskningsgången 6, 417 56 Gothenburg, Sweden

1 Introduction

Maritime education and training (MET) serve as an example of a domain that has undergone significant transformations over the previous decade: from a system of apprenticeship onboard ships to an academic education. Today, the global trend is to provide an education that provides specific competence outcomes with generic academic skills, culminating in the conferring of both maritime certificates as well as academic degrees (Manuel 2017). Along with professionalization comes standardization. In MET, the *International Convention on Standards of Training, Certification, and Watchkeeping for Seafarers* (STCW) sets out specific standards for skills and competencies that are of importance to the maritime industry (Ghosh 2017). One way of assessing these competence standards is through observable competence criteria during task-based simulator exercises in order to test if students meet the criteria for obtaining maritime certificates (Manuel 2017). The challenge now, as emphasized in numerous studies, is to ensure that simulator-based competence tests are conducted in a manner that ensures validity, reliability, and fairness of assessments (Gekara et al. 2011; Ghosh et al. 2017; Øvergård et al. 2017; da Conceição et al. 2017). The implications of this are debated in research on assessment in MET.

For example, Emad and Roth (2008, 2009) argue that the implementation of simulator-based assessment in compliance with the STCW has changed the learning objectives in MET. Instead of focus on work-relevant tasks, training became focused on passing competence test in compliance with the Convention. Emad and Roth (2008, 2009) argue that this change in learning objectives has led to the use of assessment methods that fail to measure the professional skills that enable students to put to use their competence gained from MET in workplace context onboard ships. On the other hand, Gekara et al. (2011) argue that simulator-based assessment has a distinct focus on maritime operations and professional aspects of navigation. Such aspects include maintaining vessel course and speed, safe distance from other vessels, and the required draft. Gekara et al. (2011) argue that this focus is problematic because it prevents from providing training and assessing the so-called higher cognitive skills, such as comprehension, application, analysis, synthesis, and evaluation, which is also emphasized in STCW. However, a view of maritime operations as simplistic as this might be misleading, since conducting maritime operations in a proficient and professional manner requires advanced knowledge and a high level of cognitive effort (Sellberg and Lundin 2018). Instead of questioning the validity of assessments, a recent study found that the reliability, consistency, and fairness of the assessments made through simulator-based competency tests in MET could and should be questioned (Sellberg et al. 2019). The reason is that a close relationship between instruction and assessment, resembling the corrective work that instructors perform during simulator-based training, was found to also take place during simulator-based competency tests.

With this background, the aim of this study is to investigate MET instructors' perspectives on dynamic assessment practices in simulator-based competency tests (cf. Marsh and Mitchell 2014). The study focuses on the following research questions: (a) What kind of pedagogical problems do instructors notice when viewing video materials from simulator-based competency tests? (b) How do instructors explain the problems identified in the video materials from simulator-based competency tests? In order to answer these questions, five different video-recorded episodes from authentic

instructor-student interaction during competence test were subjected to interaction analyses during *data sessions* (cf. Hindmarsh and Tutt 2012) with a group of researchers from the field of education). In the next step, the video-recorded episodes served as a basis for *video-stimulated recall* (cf. Calderhead 1981) with two focus groups with navigation instructors. The focus group discussions were video-recorded, transcribed, and subjected to discourse analysis in order to investigate how instructors approach interactions with students during simulator-based competency tests and their explanations regarding the interactions that unfold. In this manner, the adopted approach aims to provide perspectives of the phenomena under study, exploring the perspectives of both educational researchers and navigation instructors on the assessment practice. The MET institution that provides the context for this study offers education across three bachelor programs: master mariner, marine engineering, and shipping and logistics. In addition, the institution also offers master's programs and has a research school. Research at the institution is centered around different themes: maritime human factors, maritime environmental sciences, and maritime technologies. A fourth theme, maritime operations and management, is identified but focuses on teaching maritime sciences rather than researching these topics. Since 2013, the author has conducted research in this context, exploring maritime training and assessment in simulated environments through fieldwork and detailed interaction analysis of video-recorded training practices (Sellberg 2018). This study builds on this research by analyzing how simulator-based competence tests are performed in practice and what pedagogical problems and dilemmas occur during these assessments (Sellberg et al. 2019).

2 The empirical case

In Section 2.1., the process of conducting video-stimulated recall in ethnographic research is described further, going back and forth between empirical work and theoretical analysis of the assessment practice under study. In Sections 2.2. and 2.3., the setting under study and the video materials used for stimulated recall are described.

2.1 Data sessions and video-stimulated recall

Prior to the instructors' viewing of the data, 22 episodes of asking an oral exam question were selected for further analysis since the question was asked in an inconsistent way, and since the students' in the data displayed troubles in understanding the question. The episodes were then transcribed, using transcribing software InqScribe to keep a connection between data materials and transcripts. After preliminary analysis and selection, ten episodes were selected for detailed, multimodal transcriptions and further analysis (cf. Heath et al. 2010). In the next step, three episodes of the videos used for video-stimulated recall were selected for collaborative interaction analysis in a data session with researchers who were part of *the Network for the Analysis of Interaction and Learning* (LinCS-NAIL). The data session in NAIL, similar to the analytical method described in Hindmarsh and Tutt (2012), facilitates collaborative investigations of brief instances of video-recorded data in the tradition of ethnomethodology and conversation analysis. During the data sessions, multiple viewings of the

same episodes occur, supported by detailed transcripts of the interactions seen in the videos. The participating researchers then take turns on commenting on the data between viewings. By conducting repeated, detailed, and collaborative analysis of video data, the participating researchers are offered opportunities to explore tentative analyses and receive new perspectives and immediate feedback from colleagues on their empirical data (Hindmarsh and Tutt 2012). In the first data session, five senior members from the network and two visiting researchers contributed to the analysis. In addition, there were three follow-up data sessions between two or three researchers from NAIL, exploring a larger corpus of episodes from the video-recorded competence tests and, thereby, advancing the analysis.

The utilization of video in ethnography has not only opened up an avenue for close and collaborative interaction analyses of video data but has also opened up possibilities for teachers to reflect on their own teaching practice (Marsh and Mitchell 2014). Video-stimulated recall is suitable when investigating how teachers approach interactions with students in different learning activities and their own reasoning regarding the interactions that unfold (Calderhead 1981). As indicated by Marsh and Mitchell (2014), such explanations involve several levels of complexity—that is, identification and description of teaching and learning strategies, explanations that link the observations to professional knowledge, and the use of professional knowledge in order to reason around the consequences for learning. Moreover, video-stimulated recall is a technique that reveals the pedagogical beliefs and implicit theories of teachers, thereby allowing for the elicitation of knowledge-in-action (Vesterinen et al. 2010). This has proven valuable for identifying bias in teaching practices—for example, in terms of gender, as indicated in Consuegra et al. (2016). In addition, video-stimulated recall is acknowledged as a valuable tool for the professional development of teachers, providing the means for learning to identify aspects of classroom interactions (Marsh and Mitchell 2014) and for developing professional reflection-in-action through collaborative reflections on teaching practices (Geiger et al. 2016) (Fig. 1).

Video-stimulated recall has been described as a method that can be implemented for a variety of different purposes (Vesterinen et al. 2010). The approach adopted in this study aims to produce adequate descriptions of the phenomena under study by

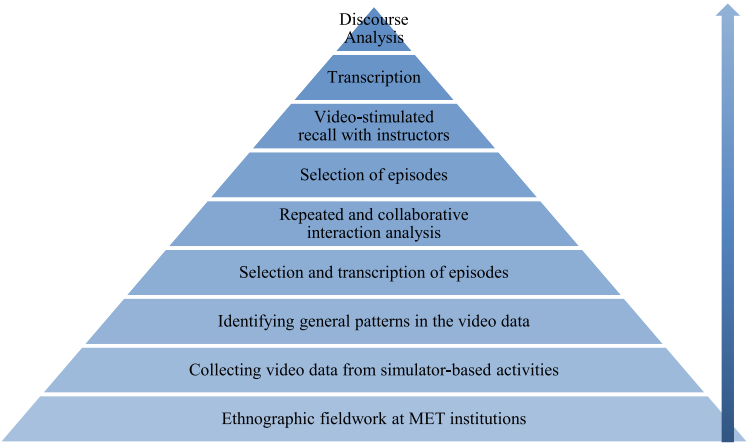

Fig. 1 An empirically driven, "bottom-up" research process

exploring the perspectives of both educational researchers and navigation instructors with regard to the assessment practice. For the current study, video records from authentic instructor-student interactions during simulator-based competence tests serve as a means for video-stimulated recall. Five simulator instructors participated in the first focus group and six simulator instructors in the second. Both focus groups were led by the same researcher. During the focus groups, the five different episodes of video-recorded data along with transcriptions of the talk served as the basis for discussions.

2.1.1 Focus group participants

The study includes 11 of the 13 instructors who teach navigation courses at the simulator center under study. Before conducting the focus groups, all participants were asked about their professional experience of working in the maritime industry, as well as their teaching experience and academic merits. The results show that the instructors in the navigation group, for the most part, have extensive experience as master mariners, varying between 1 and 30 years at sea (mean = 11.5 and median = 10). Moreover, several of the instructors have specialized navigation experience—for example, piloting or navigation in polar areas or experience from being in master positions. Moreover, the instructors also, on most part, have extensive experience as nautical teachers, with between 4 and 22 years of teaching practice (mean = 10 and median = 10). In contrast to the instructors' professional experience in the industry and as teachers, the academic merits of the group are low for an academic faculty. Four of the instructors have a master mariners bachelor's degree, six have a master of science degree in maritime management or related fields, and one has a licentiate degree and is pursuing a doctorate in maritime human factors. In the Swedish higher education system, it is also recommended for teaching staff to undertake a pedagogical course package that awards them 15 higher education credits (ECTS). Three of the instructors had taken these fifteen ECTS (or more) and five had taken single courses related to teaching and learning in higher education or industry courses aimed at MET institutions. Moreover, two of the instructors had not undergone any pedagogical courses at all, and a senior member of the group had been in the teaching position of train-the-trainer courses rather than in a student position.

2.2 Setting

The video materials used for stimulated recall were collected in a navigation course during the second year of the master mariner program (Sellberg 2018). During simulator-based training, students train to interpret and apply the *Convention on the International Regulations for Preventing Collisions at Sea* (COLREG) in various traffic situations and under different weather conditions. The results from their competence tests serve as documentation used by the Swedish Transport Agency (*swe*. Transportstyrelsen) to authorize certificates on using semi-automated technologies for navigation—that is, Radar and *Automatic Radar Plotting Aid* (ARPA) in accordance with STCW standards. Performance during the test is assessed based on several aspects. First, all overall assessment of the students' performance is based on observable actions taken during the scenario. Hence, one or two examiners, who were also instructors during the course, usually monitor the students' work in the simulator from an adjacent

instructor's room. In the instructor's room, several computer screens display different aspects of the activities on five different aspects: instrument settings, video surveillance of students' work on the bridge, and how the students perceive the marine environment through a visual look-out. The examiners also have an overall view of the scenario through a screen that displays the actions of each vessel from a bird's-eye perspective. Second, the examiner conducts observations on the bridge and asks the students various questions in order to assess the setting of different instruments. For this purpose, an assessment sheet is used to rate the students based on their work setup. Third, the examiner interviews the students on the bridge. The main purpose of these interviews is to assess the students' understanding of the use of ARPA instruments and the traffic situation at hand. As a basis for conducting interviews, the examiners use a second assessment sheet as an interview guide and to rate the students' answers.

The question selected for analysis is related to the formula for calculating rate of turn (ROT), turning radius, and speed in maneuvering a vessel. The question is formulated in the following manner in the assessment sheet: "Ask for turn indicators (3) and for the relation between rate of turn, turn ratio and speed." However, in this study, the focus is on the second portion of the questions. The students are expected to answer by accounting for the main rule of thumb: If speed is increased, the ROT is reduced if using the same turn rate that was used at the lower speed. If the speed is increased and the ROT needs to remain constant, the turn rate must be increased; otherwise, the ROT decreases. Similarly, if the speed is held constant, ROT increases if the turn rate is increased. The question is graded between 0 and 2, depending on the student's ability to identify the turn indicators and demonstrate his/her ability to explain how speed and turn rate affects the ROT.

2.3 The video material

The different episodes used as a basis for video-stimulated recall display different ways of asking questions and different ways for students to understand the question and produce an answer; results from the interaction analysis found that the instructor-student dialogues under scrutiny display instructional support in various ways. In episodes 1 and 3, the instructor's method of filling in parts of the answer in order to elicit a correct answer provides instructional support for the students (cf. Ghahari and Nejadgholamali 2019). In episodes 2 and 4, the instructor's method of grounding the question in the instruments used for navigation offers a more concrete problem formulation for the students that can make use of their navigational skills to answer the questions (cf. Säljö 2019). This is partially also the case in e5, although in this episode, the initiating question is vague with respect to the answer that the instructor is looking for. Moreover, the instructor offers encouraging words but few cues for a student who may be attempting to remember the correct answer. Overall, the instructors in these episodes, in different ways and with varying results, provide scaffolds for the student to show what they know during simulation-based competence tests (cf. Kang et al. 2014). In this manner, the assessment practice bears characteristics of being dynamic—an approach to assessment that emphasize a dialectic integration between instruction and assessments to promote learning rather than a focus on consistent rating (see, e.g., Lantolf 2009). However, the close relationship between instruction and assessment might come in conflict with reliability, consistency, and fairness of the

assessments made through simulator-based competency tests (Sellberg et al. 2019). In the next section, the instructors' viewing of the video data and their noticing of and reasoning regarding problematic situations and pedagogical dilemmas is explored.

3 Analysis and results

When initiating the focus groups, the instructors were asked to primarily discuss the students' displayed understanding of the oral question asked by the instructors in the videos and how they would assess the different answers produced by the students. The aim was to gain insight into their assessment practice by understanding how they reason in an assessment situation and how they would reason with regard to the scaffolding that was taking place during assessments. However, the discussion came to focus on several important aspects of producing an answer to a question—that is, how the question is formulated, how the dialogue unfolds, and what constitutes a correct answer, as well as implications for the consistency and fairness of the assessment. These discussions were minimally moderated, as they came to focus on such highly relevant aspects of assessments naturally. Overall, the focus groups generated 2 h of video-recorded material, which was transcribed in full by a transcription service. These transcripts focused on verbal utterances between participants. In the next step, the data corpus was analyzed using a discourse analytical approach. Discourse analysis in ethnographic research contributes to analyses of meaning beyond the words themselves, thereby exploring discourse as a situated, historic, and contingent practice (Atkinson et al. 2011). In this study, this implies that the respondents' answers are analyzed in relation to the larger context; an integral aspect of a MET institution is to uphold academic standards in line with higher education policies as well as meeting the demands of the maritime industry. During the analytical process, sections of the transcribed discussions were color coded based on topic, containing the following categories: validity, reliability, assessing the assessor, assessing the student, ideas for improvement, and challenges for improvement. While all categories are interesting, discussions on validity, reliability, and the assessment of students were selected for closer analysis. At this time, data materials from the stimulated recall were revisited in order to check the accuracy of transcriptions and to fill in previously unheard utterances where possible. Also, in accordance with ethical purposes, all participants were anonymized and ascribed a respondent number (i.e., R1, R2, and so on). All the names mentioned in their conversation were replaced with pseudonyms.

3.1 What constitutes a valid question?

In this section, discussions oriented toward the *validity* of an instructor's question in terms of the professional proficiency required of a master mariner are in focus. Here, the purpose of the question "What is the relation between ROT, turn ratio, and speed" asked by the instructor teaching the course from where data was collected (episode 1) was considered a rather practical one. R1 explains that his aim is to ascertain that a student can look at the turn rate indicator and predict the ROT. The question is further stated in the context of being able to take control over a turn from autopilot (R1) as well as related to the proficient use of the trail function in ARPA (R2). This line of

questioning is well in line with the learning objectives of the course. However, another respondent (R3) argues that this question is far from practical; rather, it is a matter of putting abstract things in correlation with one anther other, something, that does not necessarily mean that one can do this in practice:

One could answer the question without being able to perform the turn. So, it's like, it's not a skill that is being assessed in this case, it's another type of learning objective in that case… and on a rather high level of course. To put things in correlation to each other. And that is, certain abstract things, in correlation to each other. (R3).

Another respondent is continuing on this argument, describing the limitations when using higher cognitive functions versus rule-based learning during stressful situations for his colleagues:

If they should be able to use knowledge in situ, then they'll need knowledge that is more than higher cognitive functions, it has to be rule based and build on experience in order for them to be able to handle stress—that's why I think this kind of skill is more important here. (R4).

The explanation is related to current research in cognitive neuroscience, indicating how stress promotes a rapid shift in neural functions, from higher-level cognitive systems localized in the hippocampus and the prefrontal cortex in the human brain toward the habit systems involving the amygdala and the dorsal striatum (Vogel et al. 2016). Hence, in this instance, it is evident how the respondent brings his theoretical knowledge on human cognition into the discussion. The discussion displays how tensions between vocational and academic approaches to MET become practical concerns for the instructors, pondering complex questions regarding whether to educate mariners that can perform reflection-in-action at a high level of theoretical abstraction, or training mariners that can routinely and safely perform navigational operations (cf. Manuel 2017). Subsequently, after viewing episode 2, R4 elaborates on what the theoretical term "in situ"—briefly mentioned in the previous quotation—implies in such situations.

Well, here the question is more contextualized. This is… here there are things in the material world that help the students to understand. The compass spins and points, arrows n' things that this person works with. This is indeed situated cognition. And if you remove that when you… in the other question, then one removed all those supports. So, it is… I think, if you ask a question in a certain way, you will be given an answer in a certain way. These two questions are probably measuring two different sorts of knowing. (R4).

Here, R4 adopts a theoretical perspective on the observed interactions, using situated theories of cognition in his explanation. As indicated by Säljö (2019), human cognitive practices generally occur in interaction with other humans, as well as through the use of material artifacts, such as documents, computers, or—in this case—a turn rate indicator. Hence, the explanation by R4 is in line with the notion of *hybrid minds*—that is, a mind that both relies on and is shaped by its embeddedness in practices that are both social and material in nature (Säljö 2017). Following this view, knowledge or expertise must be capable of utilizing both co-workers and material artifacts to perform work in an efficient manner. Since MET has undergone significant transformations during the recent decades: from a system of apprenticeship to an academic education, the time spent learning the work practice at sea is partially replaced by training in higher education contexts. At the core, training to become a master mariner is still a matter

of learning to become part of a bridge team and working with a set of navigational tools. The situated and sociocultural perspectives on simulator-based competence tests highlight how knowledge or expertise must enable an individual to utilize both material artifacts and the knowledge and expertise of co-workers to perform work in a proficient manner. However, during the competence tests in this course, students operated independently, taking on the role of officer-of-the-watch without support from either the master in-command or a lookout. The practice of conducting individual tests has a longstanding tradition in academia, aiming to ensure that it is the knowledge of an individual that is being measured and that the grades are fair. However, this practice comes into conflict with the competence standards of MET, which aim to provide authentic assessments—that is, to assess the students' performance of work-related tasks at the standards that are required onboard ships (Ghosh et al. 2014).

Further, R4 touches upon a distinction of different kinds of knowledge but does not elaborate further on this matter in the above quotation. However, the difference between *knowing that* and *knowing how* is evident throughout the discussions (Ryle 1945). Furthermore, the relationship between *knowing that* and *knowing how* is the subject of a longstanding debate in both philosophy and educational research, since the two can be seen as either separate from each other or intricately intertwined. During this discussion, aspects that are related to this cause a lively debate. For example, going back to the quotation from R3, the question is considered as one that is assessing *knowing that* rather than *knowing how*, and the quotations from R4 indicate a perspective in favor of assessing *knowing how*, which is continuously argued. On the other hand, R1 argues in favor of the more theoretical question, indicating the need to possess generalized knowledge in order to work on different vessels as well as the absolute need to have knowledge of the theory behind how an autopilot calculates a turn in case of technical failure. While both sides make compelling arguments for their standpoints, the discussion reveals the complexity of formulating exam questions to measure the right aspect and at the appropriate level of the students' current understanding. In this teaching task, the instructors must utilize both their subject matter knowledge of navigation as well as their pedagogical knowledge of teaching, learning, and knowing.

3.2 Assessing students

During discussions on the displayed understanding of the students in different episodes, a tendency where the instructors occasionally assessed behavior and self-confidence rather than the accuracy of their answers became visible (cf. Timmermans et al. 2016). In the first focus group, the relationship between self-confidence, correctness, and unfolding dialogues was also noticed in a discussion in episode 3. However, it is interesting that the respondents were initially not able to identify the scaffolding that was taking place in the episode, where the student displayed a greater level of confidence: "Here it went well… The student knew… He could answer…." (R8). Another respondent agrees with this, filling in "but this student seems pretty confident" (R9). However, R1 challenges this view, saying "but first he gets it wrong, doesn't he?" In response to this disagreement, the interviewer replays the episode. After viewing episode 3 again, R8 noticed and communicated how the instructor elicited the correct answer by providing one of the variables for the student; moreover, R3 indicates the similarity to the instructional scaffolding in episode 1.

In focus group 2, the assessments displayed a clearer focus on assessing students' behavior and perceived character than the displayed understanding in their produced answers. For example, in a discussion on the student in episode 2, the student is described as a "super student" by R5 and "a good guy and interested" by R2. R2 continues his reasoning, saying that even if the student thinks out loud and only comes halfway to a correct answer, he is still displaying some kind of confidence in being a "good student" and "sort of on the same wavelength" as the instructor. In contrast, a student might also be considered less knowledgeable if answers are delivered in an uncertain manner. For example, the student in episode 1 who delivered a fairly good answer but in a hesitant manner was assessed in the following manner by respondents in the second focus group. "Well, for me he didn't give an especially good impression. It's a weak individual. He needs to rehearse" (R6). Another respondent in the focus group problematizes such assessments subsequently during the discussions with the following statement:

Immediately one tries to assess how good the student is. Student 1 is weak. Student 2 is great, Student 3 is kind of weak too, huh? But it can be that this is an effect of how the dialogue unfolds up until the moment we can see here now. (R7).

These findings are in line with previous studies on assessment. For example, Hirsh (2012) show that teachers tend to assess students' behavior and perceived character rather than their displayed knowledge. In fact, grading is found to be problematic, as it might be difficult to separate between students' achievements and their self-confidence as well as their performance and behavior in the classroom, or their personal bond with the teacher (Klapp Lekholm and Cliffordson 2009). On the other hand, students might be well aware of the importance of being perceived as "good students," using deliberate strategies for forming relations with teachers and playing the classroom game (Ball et al. 2012). This involves a highly social competence, that is, to skillfully decode and act upon a rather implicit curriculum (Sivenbring 2019).

What is positive in the focus group discussions that unfold after focused and repeated viewings of the video material is how the respondents collaboratively work to identify their possible biases and bring such problems up for critical discussions. As indicated in research on the use of video for analytical purposes, the possibilities for repeated viewings are said to be the analyst's best friend, since "interactions are too rich and dense to catch the details of their organization first time around" (Hindmarsh and Tutt 2012, p. 61) Moreover, it is shown how noticing pedagogical problems and bringing them up for discussion opens up for others to also take notice, reflect, and discuss such issues. The problems noted in the quoted instances of conversation led the respondents to discuss their own expectations of different students that they are training and whether or not it is appropriate for an instructor to also assess the students that they trained during a course. Moreover, the results indicate how the use of video recordings of competence tests has the potential for reducing bias in assessments by offering possibilities for repeated and collaborative assessments.

3.3 Matters of reliability, equal treatment, and fairness

In MET, as in all academic institutions, instructors and assessors are expected to promote reliable and fair testing in line with national higher educational policies, highlighting students' right to equal treatment and legal certainty. While the practice

of scaffolding during assessment enables learning-specific skills, scaffolding performed during competency tests for maritime certificates raises questions regarding the reliability of assessments as well as of equal treatment and legal certainty for MET students (Sellberg et al. 2019). Previous research emphasizes that assessment methods in MET have been developed and implemented in an ad hoc manner, i.e., not sufficiently grounded in empirical results or proven experience, and are leading to inaccurate and inconsistent assessment practices that cause more harm than good if perceived as reliable measures of professional performance (Ernstsen and Nazir 2018). During focus group discussions with instructors, it becomes evident that they themselves are highly critical of the reliability and fairness of simulator-based competence tests:

I struggle with this issue a lot. I'm spending a lot of time in the simulator and do a lot of examinations and I feel it. It's not fair. I'm not doing a fair assessment. It's so many parameters that affect… just that they are doing a scenario… someone has a small… they don't do the exact same scenarios. (R1).

In the simulator-based scenarios, the students are beginning from different positions to cross the Dover Strait. Consequently, this scenario design provides students with different premises for performing the task. Moreover, during the tests, five students are navigating the same scenario, sharing the virtual world from different bridges. While simulators are perceived as offering a controlled environment, there are still aspects of navigating in traffic that are difficult to control, as students' interactions with each other in the scenario might unfold in unpredictable ways. Hence, in simulator-based competency tests, a scenario design where students begin their crossing from the same position and navigate in a simulated scenario where all meeting traffic is predesigned might reduce inconsistencies during simulator-based competency tests.

Another problem brought up during discussions is how to assess competency in a reliable manner and the role of predesigned assessment sheets when conducting assessments:

What's important is what the objective is, what are the learning objectives? How do we assess them and how do we follow up on our assessments, so to say? And then… how do we handle the assessment in relation to criteria, where do we draw the line of what is okay, this is on the verge or this is really good? It's like we… one feels that we need more of a model to work from. (R5).

Further, R7 indicates that a predesigned assessment sheet, even when it is thoroughly designed, is difficult to implement in a reliable manner:

This doesn't concern questions it's based on observations during scenarios. And it's hard to see what we should put focus on really… What is it that is good? And it's very much relying on feelings, like I feel that this is good and that's no good.

Here, R7 refers to a course on bridge teamwork in offshore operations, where an assessment sheet has been designed and implemented in line with the Nautical Institute's standards and research on how to assess non-technical skills (cf. Conceição et al. 2017). In keeping with recommendations on how to use such assessment models, only observable behavior is assessed in the course, which reduces the kind of inconsistencies evident in this data set, where the dialogue after asking oral questions unfolds in eclectic ways. However, as evident through R7's concerns, implementing such assessment models is not easy, thereby revealing how a model in itself is no warrant for reliability in assessments. As indicated in research from aviation, where such assessment models have a long tradition, interrater reliability between examiners has been

shown to be moderate to low (e.g., Weber et al. 2013; Roth 2015). One reason for this is that even if the assessment models break down observations into separate measurable entities of skills, flight examiners based their assessments on a large number of observations put together into a holistic event, even when using rating scales (Roth 2015). This may be because technical skills and non-technical skills are difficult to isolate and target during assessment, and because skillful behavior requires to be understood as a whole (Sellberg and Lundin 2018). In this context, the respondents in this study describe the tension between skillful professional practice and the quest for deconstructing and measuring. Their reasoning can be summed up in a quote from one of the instructors:

I think that… On one hand measurability… There is justice in measurability. And on the other hand, we have the instructors' decisions. Did this go well or didn't it go well? I understand the eagerness to be fair and the pathos for justness, it's great in many respects. But at the same time, it's like… You have a certain kind of knowledge, a sense where you can assess the student. We need to make this as fair as possible, but we need to keep these aspects of being skillful, being a human and having this sense of… It's not really in line with how science is constructed in a technical university, but it's still important. (R4).

As indicated in (Sellberg 2018) the endeavor to increase reliability by deconstructing maritime operations into a set of discrete entities might be leading MET in the wrong direction. The tension between scientific rationality and professional reflection-in-action has been extensively debated, often in relation to the work of Schön (1983, 1987) and occasionally in favor of one of the other. In fact, Schön's (1987) approach makes arguments in favor of the practitioners' use of scientific methods for further developing their reflection-in-action and reflection-on-action. This approach might be the most fruitful for developing training and assessment in MET—that is, further developing instructors' professional reflective practicum, both as mariners and as teachers (cf. Schön 1987).

4 Discussion

This study shows how longstanding theoretical and philosophical debates between *knowing that* and *knowing how* and between scientific rationality and the art of skillful practice manifest as practical dilemmas in MET instructors' teaching tasks. Hence, tensions in MET between academic procedures and a vocational focus require further attention in research on assessment in MET. Moreover, the academization of MET has changed not only the procedures for assessment but also the role of teachers: from masters guiding apprentices to instructors, lecturers, and professors teaching students to help them obtain academic degrees. While professionalization is related to enhancing the status of work, the overall aim is to improve the quality of the members and the services provided by their work (e.g., Hoyle 2001). In practice, this implies that instructors require professionalized knowledge on the subject matter of their teaching in maritime courses as well as a professionalized understanding of their own teaching practice. However, during fieldwork in MET institutions, another discourse on university teaching regularly surfaces—that is, an attitude that the pedagogical knowledge and didactical skills of university teachers is common sense understanding that everyone

holds or a talent that comes naturally for some but not for others. This attitude is certainly not unique for MET instructors, as reported by Maphosa and Mudzielwana (2014, p 65):

Discourse in university often revolves around that teaching is common sense business and that anyone can teach. University lecturers as experts in their disciplines often feel they are able to teach and it is up to students to learn.

Such attitudes might explain why the instructors in this study had varying degrees of pedagogical training, as well as why some of them had worked for several years without given or taking the opportunity to go through basic courses on teaching and learning in higher education. This is problematic, neglecting the pedagogical complexities of university teachers' responsibilities—that is, curriculum organization, teaching, and assessment. In contrast, a professionalized teaching practice can be understood as a process through which teachers advance their levels of professional competence throughout their careers (Fernández 2013). This process includes a learning period when entering the academic teaching profession—teaching and learning through higher education courses and the teachers' self-directed learning. In this process, common sense knowledge on teaching and learning serve as a starting point for teachers to begin developing their professional competence through reflection on their own beliefs and analyzing them in the context of formal theories of teaching and learning (Tatto 2019). This, in turn, implies that the teacher recognizes areas in their own teaching practice that require improvement and works in a structured and scientific manner to advance their teaching, incorporating current research on their subject matter as well as teaching and learning into their teaching practice (Fernández 2013). In this manner, the professional development of university teachers is described as an evolution toward professionalism, characterized by the ability to analyze, critically judge, and develop the processes involved in teaching by relating professional experience with theories of learning. However, it is important to highlight that even if these activities are described as self-regulated practices, as in Fernández (2013), such intellectual work might be difficult to achieve individually. Rather, these activities must be viewed as becoming part of an academic culture that provides teaching staff with opportunities, resources, and infrastructures for continuous, collaborative, and structured pedagogical development.

5 Conclusion

A current challenge in MET is to ensure that simulator-based competence tests are conducted in a manner that ensures validity, reliability, and fairness of assessments. The results from this study provide perspectives on these issues, revealing the dynamic nature of assessing competence as well as how instructors participating in focus groups identified and critically discussed a variety of pedagogical problems and dilemmas related to these dynamics. These discussions oriented toward longstanding issues of validity as well as questions of consistency and fairness of the assessments made; they also reveal precisely *how* and *why* conducting valid and reliable assessments of competence is a complex educational matter. The results highlight the following key findings in regard to conducting high-stake assessments for maritime certificates:

- The possibilities to strengthen reliability as well as the equal treatment and legal certainty toward students if basing assessment on video records of competence tests rather than during simulator tests.
- How holding collaborative and regular discussions on assessment situations might be a fruitful approach to reduce individual bias in grading.
- How thoughtful scenario designs might reduce inconsistencies in testing.
- The role and importance of navigation instructors' professional knowledge and theoretical understanding of their practice, both as mariners and as university teachers, when performing simulator-based competency tests.

Previous studies on strengthening validity, reliability, and fairness of simulator-based competence tests often aim at reducing the subjective impact of the instructor/assessor by providing the means for objective measures of professional competence, .e.g., through computer-assisted assessment tools (e.g., Ernstsen and Nazir 2020) or through models of assessing non-technical skills (da Conceição et al. 2017). Results from the current study take on a different perspective and suggest strengthening the pedagogical competence of the instructor/assessor as a possible way forward. One of the implications of this research calls for allocating resources to support and advance the pedagogical development of MET instructors, thereby highlighting the importance of matching their professional knowledge of maritime work with expertise in teaching, learning, and assessment. However, further studies on simulator-based assessment of competence in MET are needed to explore the effects of such interventions.

Acknowledgments Open access funding provided by University of Gothenburg. I would also like to acknowledge my colleagues in LinCS-NAIL for the analytical contributions of the video material used for stimulated recall. In this work, Roger Säljö deserves special recognition for his continuous analytical interest and discussions on the material. I am also grateful to the navigation instructors who continue to share their teaching practice for me to analyze. Further, I thank Olle Lindmark for support in organizing and conducting the focus groups.

Funding information This research is funded by FORTE (Swedish Research Council for Health, Working Life, and Welfare) project no: 2018-01198.

References

Atkinson D, Okada H, Talmy S (2011) Ethnography and discourse analysis. In: Hyland K, Paltridge B (eds) Continuum companion to discourse analysis. Bloomsbury, London, pp 85–100

Ball JS, Maguire M, Braun A (2012) How schools do policy: policy enactments in secondary schools. Taylor and Francis, Hoboken

Calderhead J (1981) Stimulated recall: a method for research on teaching. Brit J Educ Psychol 51:211–217. https://doi.org/10.1111/j.2044-8279.1981.tb02474.x

Consuegra E, Engels N, Willegems V (2016) Using video-stimulated recall to investigate teacher awareness of explicit and implicit gendered thoughts on classroom interactions. Teach Teach 22:683–699. https://doi.org/10.1080/13540602.2016.1158958

da Conceição VP, Basso JC, Lopes CF, Dahlman J (2017) Development of a behavioral marker system for rating cadet's non-technical skills. TransNav: Int J Mar Navig Saf Sea Transp. https://doi.org/10.12716/1001.11.02.07

Emad G, Roth WM (2008) Contradictions in the practices of training for and assessment of competency. Education and Training 50(3):260–272

Emad G, Roth WM (2009) Policy as boundary object: a new way to look at educational policy design and implementation. Vocations and Learning 2(1):19–35

Ernstsen J, Nazir S (2018) Consistency in the development of performance assessment methods in the maritime domain. WMU J Marit Aff 17:71–90. https://doi.org/10.1007/s13437-018-0136-5

Ernstsen, Nazir S (2020) Performance assessment in full-scale simulators–a case of maritime pilotage operations. Safe Sci 129:104775. https://doi.org/10.1016/j.ssci.2020.104775

Fernández JT (2013) Professionalization of teaching in universities: implications from a training perspective. Int J Educ Technol High Educ 10. https://doi.org/10.7238/rusc.v10i1.1471

Geiger V, Muir T, Lamb J (2016) Video-stimulated recall as a catalyst for teacher professional learning. J Math Teacher Educ 19:457–475. https://doi.org/10.1007/s10857-015-9306-y

Gekara VO, Bloor M, Sampson H (2011) Computer-based assessment in safety-critical industries: the case of shipping. J Vocat Educ Train 63:87–100. https://doi.org/10.1080/13636820.2010.536850

Ghahari S, Nejadgholamali A (2019) Instructed assessment and assessed instruction: a review of dynamic assessment and ways forward. Educ Psychol Pract 35:384–394. https://doi.org/10.1080/02667363.2019.1617113

Ghosh S (2017) Can authentic assessment find its place in seafarer education and training? Aust J Marit Ocean Aff 9:213–226. https://doi.org/10.1080/18366503.2017.1320828

Ghosh S, Bowles M, Ranmuthugala D, Brooks B (2014) Reviewing seafarer assessment methods to determine the need for authentic assessment. Aust J Marit Ocean Aff 6(1):49–63. https://doi.org/10.1080/18366503.2014.888133

Ghosh S, Bowles M, Ranmuthugala D, Brooks B (2017) Improving the validity and reliability of authentic assessment in seafarer education and training: a conceptual and practical framework to enhance resulting assessment outcomes. WMU J Marit Aff 16:455–472. https://doi.org/10.1007/s13437-017-0129-9

Heath C, Hindmarsh J, Luff P (2010) Video in qualitative research. Analysing social interaction in everyday life, London, SAGE

Hindmarsh J, Tutt D (2012) Video in analytic practice. In: Pink S (ed) Advances in visual methodology. SAGE, New Dehli, pp 57–73

Hirsh Å (2012) The individual education plan: a gendered assessment practice? Assess Educ: Princ, Policy Pract 19:469–485. https://doi.org/10.1080/0969594X.2012.694587

Hoyle E (2001) Teaching as profession. In: Smelser NJ, Baltes PB (eds) International encyclopedia of the social & behavioral sciences. Pergamon Press, Oxford, pp 15472–15476

Kang H, Thompson J, Windschitl M (2014) Creating opportunities for students to show what they know: the role of scaffolding in assessment tasks. Sci Ed 98:674–704. https://doi.org/10.1002/sce.21123

Klapp Lekholm A, Cliffordson C (2009) Effects of student characteristics on grades in compulsory school. Educ Res Eval 15:1–23. https://doi.org/10.1080/13803610802470425

Lantolf JP (2009) Dynamic assessment: the dialectic integration of instruction and assessment. Lang Teach 42:355–368. https://doi.org/10.1017/S0261444808005569

Manuel ME (2017) Vocational and academic approaches to maritime education and training (MET): trends, challenges, and opportunities. WMU J Marit Aff 16:473–483. https://doi.org/10.1007/s13437-017-0130-3

Maphosa C, Mudzielwana NP (2014) Professionalization of teaching in universities: a compelling case. Int J Educ Sci 6:65–73. https://doi.org/10.1080/09751122.2014.11890119

Marsh B, Mitchell N (2014) The role of video in teacher professional development. Teach Dev 18:403–417. https://doi.org/10.1080/13664530.2014.938106

Øvergård KI, Nazir S, Solberg A (2017) Towards aAutomated performance assessment for maritime navigation. TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation 11(2):43–48

Roth W-M (2015) Flight examiners' methods of ascertaining pilot proficiency. Int J Aviat Psychol 25:209–226. https://doi.org/10.1080/10508414.2015.1162642

Ryle G (1945) Knowing how and knowing that: the presidential address. Proc Aristot Soc 46:1–16 https://www.jstor.org/stable/4544405. Accessed 5 April 2020

Säljö R (2017) Conceptual change, materiality and hybrid minds. In: Amin AG, Levrini O (eds) Converging perspectives on conceptual change: mapping an emerging paradigm in the learning sciences. Routledge, London, pp 113–120

Säljö R (2019) Materiality, learning, and cognitive practices: artifacts as instruments of thinking. In: Cerratto Pargman T, Jahnke I (eds) Emergent practices and material conditions in learning and teaching with technologies. Springer, Cham, pp 21–32

Schön DA (1983) The reflective practitioner: how professionals think in action. Basic Books, New York

Schön DA (1987) Educating the reflective practitioner: towards a new design for teaching and learning in the profession. Jossey-Bass, San Francisco

Sellberg C (2018) Training to become a master mariner in a simulator-based environment (Doctoral dissertation, Thesis for the Degree of Doctorate of Education), University of Gothenburg, Gothenburg, Sweden.

Sellberg C, Lundin M (2018) Tasks and instructions on the simulated bridge: discourses of temporality in maritime training. Discourse Stud 20(2):289–305

Sellberg C, Lindmark O, Lundin M (2019) Certifying navigational skills: a video-based study on assessments in simulated environments. TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation 13(4):881–886

Sivenbring J (2019) Making sense and use of assessments. Scand J Ed Res 63:759–770. https://doi.org/10.1080/00313831.2018.1434827

Tatto M (2019) The influence of teacher education on teacher beliefs. Oxf Res Encycl Educ. https://oxfordre.com/education/view/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-747. Accessed 5 April 2020

Timmermans AC, de Boer H, van der Werf MP (2016) An investigation of the relationship between teachers' expectations and teachers' perceptions of student attributes. Soc Psychol Educ 19:217–240. https://doi.org/10.1007/s11218-015-9326-6

Vesterinen O, Toom A, Patrikainen S (2010) The stimulated recall method and ICTs in research on the reasoning of teachers. Int J Res Method Educ 33:183–197. https://doi.org/10.1080/1743727X.2010.484605

Vogel S, Fernández G, Joëls M, Schwabe L (2016) Cognitive adaptation under stress: a case for the mineralocorticoid receptor. Trends Cogn Sci 20:192–203. https://doi.org/10.1016/j.tics.2015.12.003

Weber D, Roth W-M, Mavin T, Dekker S (2013) Should we pursue interrater reliability or diversity? An empirical study of pilot performance assessment. Aviat Focus–J Aeronaut Sci 4(2):34–58