INTRODUCTION

A highly skilled workforce is essential in the seafaring industry to navigate through diverse operational scenarios. Successful operation depends on competent human operators where training plays a crucial part. Simulators are utilized for various seafarer training contexts where risk-free, repeated exercises are facilitated with considerably less time and cost than other traditional on-the-job training methods [1], [2]. Assessing trainee performance bears crucial significance in the overall success of maritime training whether it is classroom-based or simulator-based training [3], [4]. The conceptual coupling between specific learning outcomes and the assessment methods employed in training is also crucial from a theoretical point of view [5]. On the other hand, the employers require reliable indicators and evidences of seafarers’ competence which forms the practical need of assessment [6]. Further, to justify the growing cost of training and its impact on workplace performance, standardized assessment scales have emerged as a means of evaluation [7]. In addition, the challenges related to attending the “workplace relevance” in maritime training and assessment only adds to the complexity of practical requirements [6].

1.1 Conceptual foundations of assessment

The definition of “educational assessment” has taken many forms over the years and is still under scrutiny. In the early 1990s, educational assessment simply meant to measure the outcome of learners’ achievements relative to either their peers or based on their performance [8], [9]. There are two existing school of thoughts related to assessment of educational output namely the realist and the
relativist approaches [10]. The realists tend to measure performance with regards to criterion-based standardized scales whereas the relativists rely on expert judgement in assessing learner’s performance. Amidst the growing variation of assessment methods and confusion surrounding their application, Kraiger et al. (1993) suggested specific assessment methods to be employed to the corresponding learning outcomes, i.e., cognitive, skill-based, and affective outcomes (see Table 1).

In addition, authentic assessment is continually getting popular where assessment tasks are replicated according to the desired performance of real workplace, therefore, termed as “performance assessment” [8]. We also seek out “what constitutes a good assessment?”. Gipps (1994) noted a few key elements of good assessment practices which include:
- a range of activities and wide opportunity to perform,
- carried out in a safe environment, i.e., a normal classroom setting with ample opportunity to interact with teacher,
- a range of media to perform except written exams.

Table 1. Prescribed assessment methods for differing learning outcomes (adapted from Kraiger et al., 1993)

<table>
<thead>
<tr>
<th>Learning outcomes</th>
<th>Assessment methods</th>
<th>Measurement goals</th>
<th>Tools and method descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive outcomes:</td>
<td>Recognition and recall tests</td>
<td>Measuring declarative knowledge</td>
<td>Multiple-choice question, true-false, free recall exams</td>
</tr>
<tr>
<td>a compilation of problems</td>
<td>Power tests</td>
<td>Measuring the accuracy of knowledge</td>
<td>The number of correct answers, without any time limit</td>
</tr>
<tr>
<td>solving strategies that</td>
<td>Speed tests</td>
<td>Measuring the rate at which an individual can access</td>
<td>The number of correct answers, or the reaction time to any single item</td>
</tr>
<tr>
<td>include declarative and</td>
<td>Free sorts</td>
<td>knowledge, structure, cognitive maps, or mental models</td>
<td>Observing if trainees can physically arrange elements according to expected clusters</td>
</tr>
<tr>
<td>procedural knowledge</td>
<td>Structural assessment</td>
<td>(same as above)</td>
<td>Clustering or scoring algorithms</td>
</tr>
<tr>
<td>[11]</td>
<td>Probed protocol analysis</td>
<td>Measuring cognitive strategies and task behaviour</td>
<td>Questionnaire asking trainees specified probe questions at each task step</td>
</tr>
<tr>
<td></td>
<td></td>
<td>relative to goals</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Self-report</td>
<td>Measuring trainees’ self-awareness of knowledge gained</td>
<td>Questionnaire related to the awareness of procedural level, additional learning requirements and awareness of mistakes</td>
</tr>
<tr>
<td></td>
<td>Readiness for testing</td>
<td>Measuring the ability of learners to judge the</td>
<td>Questionnaire</td>
</tr>
<tr>
<td></td>
<td></td>
<td>future applicability of their knowledge</td>
<td></td>
</tr>
<tr>
<td>Skill-based outcomes:</td>
<td>Targeted behavioural observation</td>
<td>Measuring the speed and fluidity of performance, error</td>
<td>Observing frequency of desired behaviour, time, steps and</td>
</tr>
<tr>
<td>skill-based outcomes</td>
<td></td>
<td>rates, overall conceptualization, and trainees’ ability to generalize skills in different contexts</td>
<td>sequencing requirements to complete a task, error counts etc.</td>
</tr>
<tr>
<td>related to technical and</td>
<td>Hands-on testing</td>
<td>Measuring available cognitive resources and the</td>
<td>Observing if trainees can identify a series of correct steps and providing qualitative evaluation along the way</td>
</tr>
<tr>
<td>motor skills [5]</td>
<td></td>
<td>automaticity of performance</td>
<td>Questionnaire asking trainees about how they would perform a particular task</td>
</tr>
<tr>
<td></td>
<td>Structured situational interviews</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Secondary task performance</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interference problems</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Embedded measurement</td>
<td>(same as above)</td>
<td></td>
</tr>
<tr>
<td>Affective outcomes:</td>
<td>Self-report measures</td>
<td>Measuring trainee attitudes, mastery, perceived</td>
<td>Questionnaire asking trainees about their capability and</td>
</tr>
<tr>
<td>an internal state of</td>
<td></td>
<td>performance capability and goal setting</td>
<td>confidence in performing a task with varying difficulties</td>
</tr>
<tr>
<td>behaviour which affect</td>
<td>Free recall measures</td>
<td>Measuring complexity of goal structures</td>
<td>Questionnaire, focused interviews or think aloud protocol for trainees</td>
</tr>
<tr>
<td>learners’ attitude [11]</td>
<td></td>
<td>Measuring goal commitment</td>
<td>Observing if trainees can physically arrange elements according to expected clusters</td>
</tr>
</tbody>
</table>

The requirements of good assessment resonate with the literature related to maritime simulator training where trainees participate in a wide range of scenarios in a safe environment and in various modalities, i.e., full-mission, desktop-based, cloud and virtual reality (VR) simulators [2, 4, 12–15].

1.2 Assessment in maritime simulator training

The demand for diverse assessment methods in maritime training has arisen to meet complex learning objectives, such as ensuring pedagogical effectiveness [14] and developing practical job skills [6]. Consequently, a plethora of assessment methods have been developed and are currently utilized in maritime simulator training to address both theoretical (i.e., pedagogical) and practical (i.e., job relevant) aspects of assessment. Debriefing [1], dynamic assessment by analysing video data [14], psychophysiological evaluation [16] and differing computer-based tools [3], [17] are some of the sighted methods of assessment in the mentioned context.
However, research indicates that these assessment methods may not guarantee the acquisition of competency and the development of a competent workforce [4], [6], [18]. The lack of authenticity in current assessment practices during simulator training is considered a hindrance to the development of essential competencies and preparation for workplace challenges. Consequently, the need for "authentic assessment" has been highlighted in the literature [6], [19]–[21].

In addition, instructors’ subjectivity, bias, and uncertainty around assessment methods pose significant challenges in operationalizing efficient assessment during simulator training [3], [4]. For example, instructors may develop scenarios catering to a particular group of trainees or test students on simulators based on their own experience (e.g., cargo ship experience), which may not be suitable for other trainees on a different context (e.g., emergency on a passenger ship).

Maritime simulators provide a simulated virtual environment for education, making them ideal technology-based learning environments [22]. Thus, the same assessment challenges that exist in other technology-based learning environments also apply to maritime simulator training. Meeting the challenges of maritime simulator training requires addressing differences in learner characteristics, technical capabilities, and pedagogical design. These differences can lead to inconsistent learning outcomes. For example, variations in learners’ characteristics and simulator fidelity can result in mixed learning outcomes [23]. In addition, the misalignment of simulation practices with pedagogical theories is a recognized issue [24].

Assessing trainee performance in maritime training is a complex task with varying opinions on its objectives, especially during simulator training. Some view assessment as a means of objectively categorizing maritime trainees based on competence [3], [25], while others see it as an assistive learning instrument [15], [26], [27]. Objective assessment aims to ensure the validity and reliability of evaluation measures, but it is a fallacy to believe that learning can be accurately and reliably assessed [8]. The need for professional intersubjectivity of instructors undermines the requirements of validity and objectivity of assessment measures [15], [28]. Sadler (2005) proposed that the ideal performance assessment should focus on "standard performance" rather than "criterion-based objective assessment" focusing on validity only. There is a lack of consensus if assessment of maritime simulator trainees should purely aim for objective measures or depend on expert evaluation of instructors. Ideal assessment frameworks can be conceptualized as a method where trainee performance can be reliably assessed without compromising the expert-in-the-loop feature.

Therefore, in-depth knowledge on the operationalization of existing assessment methods in maritime simulator training is crucial to ensure that theoretical and practical aspects of evaluation are satisfied. It is also important to identify which evaluation methods are best suited for specific simulator training scenarios and their specific advantages and limitations. Such analysis would pave the way for assessment best-practices that can help maritime instructors to administer appropriate assessment methods for specific training needs, while also being aware of their benefits and pitfalls. Awareness of the assessment best-practices can also assist instructors in developing valid and reliable assessment scenarios and adapting their training approaches to achieve desired learning outcomes. Furthermore, educators would acquire enhanced comprehension regarding the optimal timing and extent of expert intervention in the evaluation process, enabling them to make informed decisions on when and how much to intervene.

1.3 The goal of this study

In this study, we examined the empirical studies that employed various assessment methods within the context of maritime simulator training. The goal was to systematically assemble differing assessment methods including their objectives, their specific advantages and limitations coupled with the in-depth analysis of their suitability with differing learning outcomes. Thus, we also shed light on the state-of-the-art of assessment methods identifying their gaps and discuss the future requirements in maritime simulator training context.

The following research questions have been formed:
- RQ1: How differing assessment methods are operationalized in maritime simulator training?
- RQ2: What are the advantages and limitations of operationalized assessment methods in maritime simulator training?

2 METHODS

To address the research questions, a systemic literature review method was adopted in this study. The following keywords were utilized for searching documents in two different databases (i.e., Web of Science and Scopus): ("maritime" or “shipping” or “seafarer*”) AND “simulator*” and “training” and “assessor”. A set of inclusion and exclusion criteria is used during the database search (see Table 2).

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peer-reviewed</td>
<td>Not related to simulator training</td>
</tr>
<tr>
<td>English articles</td>
<td>No intervention used during the study or experiment</td>
</tr>
<tr>
<td>Seafarer training related empirical studies</td>
<td></td>
</tr>
<tr>
<td>Studies operationalizing Conceptual and non-specific assessment method</td>
<td></td>
</tr>
<tr>
<td>Studies with clear learning/ training outcome (Cognitive, reports) White paper or technical</td>
<td></td>
</tr>
</tbody>
</table>

Skill-based or Affective outcomes |

The initial number of search output from two (02) separate databases resulted in a total of 147 studies including one (01) document added through snowballing. The overall literature review process followed a systemic approach as depicted in Figure 1 aligning with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA).
of simulator used, assessment measures and methods along with the associated learning outcomes. In addition, the advantages and limitations associated with each assessment method as mentioned in the studies are also included in the analysis. Another co-author separately screened the excerpts of the analysis in excel format for inter-rater reliability (see Table 3).

Table 3. Performance assessment in maritime simulator training

<table>
<thead>
<tr>
<th>SL</th>
<th>Source</th>
<th>Assessment goals</th>
<th>Simulator types</th>
<th>Assessment measures</th>
<th>Corresponding method</th>
<th>Learning outcome</th>
<th>Advantages</th>
<th>Limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[31]</td>
<td>Measuring situational awareness of navigational officers</td>
<td>Bridge simulator</td>
<td>Reaction time to failure</td>
<td>Speed test; secondary task performance</td>
<td>Cognitive, Skill-based</td>
<td>(Not listed in the study)</td>
<td>Overall time limit of the exercises. It can be argued that participants could correct the issues if given enough time.</td>
</tr>
<tr>
<td>2</td>
<td>[32]</td>
<td>Assessing electronic navigation competency</td>
<td>Bridge simulator</td>
<td>Eye tracking data for fixation duration and fixation counts</td>
<td>Targeted behavioural observation</td>
<td>Skill-based</td>
<td>Eye-tracking provides novel data, e.g., focus of attention, facilitates objective observation of trainees. Inter-rater reliability was fair for CAPA-tool compared to conventional methods in terms of teamwork performance.</td>
<td>Methodological challenges related to utilizing AHP tools for weighting performance score, CAPA tool relies heavily on human observation or interpretation, thus human bias cannot be fully eliminated, Criterion validity of the tool could not be established.</td>
</tr>
<tr>
<td>3</td>
<td>[3]</td>
<td>Assessing pilotage operation</td>
<td>Bridge simulator</td>
<td>Computer-aided Performance Assessment tool (CAPA) utilizing Analytical Hierarchical Process (AHP) and Bayesian Network for binary assessment of checklist items</td>
<td>Targeted behavioural observation</td>
<td>Cognitive, Skill-based</td>
<td>Opportunity for objective and continuous monitoring of students by eye tracking.</td>
<td>High cost of eye tracking equipment. Limitation of analysis software to analyse dynamic scenarios as ships’ motion is unstable. The bigger size and layout of bridge simulators is not optimal for the analysis of large amount of eye-tracking data.</td>
</tr>
<tr>
<td>4</td>
<td>[33]</td>
<td>Assessing situational awareness of bridge watchkeepers</td>
<td>Bridge simulator</td>
<td>Eye tracking data for fixation duration, heatmap along with debriefing</td>
<td>Targeted behavioural observation; hands-on testing</td>
<td>Skill-based</td>
<td>Time consuming; inclusion of subjective factors in the assessment such as defining optimal measures (e.g., trigger time).</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>[34]</td>
<td>Assessing visual attention during heavy lifting operation</td>
<td>Heavy lift simulator</td>
<td>Eye-tracking, briefing-debriefing, questionnaire</td>
<td>Targeted behavioural observation; hands-on testing; recognition &amp; recall test</td>
<td>Psychophysiological evaluation complements the current simulator-aided assessment. High reliability of the proposed system. (Not listed in the study)</td>
<td>(Not listed in the study)</td>
<td>Smaller sample size and lack of variation in the vessel types being used during assessment.</td>
</tr>
<tr>
<td>6</td>
<td>[16]</td>
<td>Psychophysiological (cognitive workload, stress) evaluation of seafarers</td>
<td>Bridge simulator</td>
<td>EEG for measuring heart rate variability</td>
<td>Targeted behavioural observation</td>
<td>Psychophysiological evaluation complements the current simulator-aided assessment. High reliability of the proposed system. (Not listed in the study)</td>
<td>Time consuming; inclusion of subjective factors in the assessment such as defining optimal measures (e.g., trigger time).</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>[12]</td>
<td>Assessing the effect of introducing complexity in different levels of simulator training</td>
<td>Bridge simulator</td>
<td>ECDIS data to calculate cross-track error</td>
<td>Targeted behavioural observation</td>
<td>(Not listed in the study)</td>
<td>Smaller sample size and lack of variation in the vessel types being used during assessment.</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>[35]</td>
<td>Evaluating individual risk perception with a focus on situational awareness</td>
<td>Offshore crane simulator</td>
<td>Multi-layer and multi-sensor fusion to analyse bio-metric data</td>
<td>Targeted behavioural observation; hands-on testing</td>
<td>Provides new insights into novel SA assessment methodologies</td>
<td>Lack of robustness and validity concerns with methods.</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>[36]</td>
<td>Measuring SA during an engineering-supervisory task environment in ships</td>
<td>Engine plant simulator</td>
<td>NASA-TLX for perceived SA, SA sensitivity</td>
<td>Self-report; structured situational interviews</td>
<td>Cognitive</td>
<td>Objective measure: the simulation required to be frozen several times to administer the questionnaire; subjective measure: limited in depicting the effect of different level of participants expertise during measurement; familiarity effect was higher for subsequent measures; sensitive to different workload levels; participants’ actions during the scenario was passive; the mental</td>
<td></td>
</tr>
<tr>
<td>Number</td>
<td>Method</td>
<td>Simulator/Equipment</td>
<td>Observation/Testing</td>
<td>Evaluation Method/Tool</td>
<td>Remarks</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>------------------------------------------------------------------------</td>
<td>---------------------</td>
<td>------------------------------------------</td>
<td>------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Evaluating navigational competence</td>
<td>Bridge simulator</td>
<td>Observation with assessment sheet indicating rate of turn, turn rate and speed in manoeuvring along with interviews</td>
<td>Hands-on testing; probed protocol analysis; embedded measurements</td>
<td>Skill-based, Cognitive Keeps instructor in the loop during assessment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Evaluating lifeboat launching operation</td>
<td>Lifeboat simulator</td>
<td>Questionnaire including accuracy of recall, order of actions</td>
<td>Power test; hands-on testing</td>
<td>Cognitive, Skill-based Verbal administration was possible without lowering the lifeboats</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Assessing navigational competency</td>
<td>Bridge simulator</td>
<td>Observation, interviews, video recording for COLREG deviations</td>
<td>Hands-on testing; targeted behavioural observation</td>
<td>Skill-based Possibility of instructor-in-the-loop during training, i.e., instructors intervene during training to correct mistakes and give inputs to fulfill the learning objectives</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Human error evaluation</td>
<td>Engine simulator</td>
<td>Questionnaire on different operational tasks on mimic panels</td>
<td>Targeted behavioural observation; hands-on testing</td>
<td>Skill-based Possible to measure both operational task competency and situational awareness, and can be used to select best candidates for certain operation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Evaluate and rank students based on their actions with regards to different error producing conditions</td>
<td>Engine simulator</td>
<td>Shipboard Operation Human Reliability Analysis (SOHRA) for overall trainee performance AHP, Evidential Reasoning (ER) algorithm for identified behavioural markers</td>
<td>Targeted behavioural observation; hands-on testing</td>
<td>Skill-based This method can be utilized to select best candidate for a specific operation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Assessing deck officers NTS</td>
<td>Bridge simulator</td>
<td>Computer Based Evaluation (CBE) including course, speed, overshoot angle, rudder angle, track deviation and time Observation, checklist including Rules of the Road, communication, vessel positioning, look-out, manoeuvring</td>
<td>Targeted behavioural observation; hands-on testing</td>
<td>Skill-based The effectiveness of any training methodology can be determined</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Evaluating manoeuvring performance in differing situations</td>
<td>Bridge simulator</td>
<td>航海安全与操作模拟器</td>
<td>Hands-on testing; targeted behavioural observation</td>
<td>Skill-based CBE provides additional support for simulator instructors and provides opportunity for increasing the objectivity in evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Proposing an assessment method for evaluating the required time for training and the level of navigational competency</td>
<td>Bridge simulator</td>
<td>Self-reported Situational Awareness Rating Scale (SARS) questionnaire, ECG for SA, workload</td>
<td>Self-report</td>
<td>Cognitive, Affective Provides new insights into how SA affects learning outcome during simulator training</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Assessing perceived situational awareness (SA), learning outcome and perceived realism</td>
<td>Bridge simulator</td>
<td>Self-reported Situational Awareness Rating Scale (SARS) questionnaire, ECG for SA, workload</td>
<td>Self-report</td>
<td>Cognitive, Affective Provides new insights into how SA affects learning outcome during simulator training</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

model of participants action or communication instances were not evaluated. There is scope for assessors’ subjectivity and individual bias.

The type of questionnaire favoured one type of simulator participants more than the other. As instructors provide selective correction to a few students but not all, the fairness of assessment is questioned.

(Not listed in the study)

Inadequate data for conclusive results.

Lack of clear evaluation criteria, communication and individual situation-awareness aspects are difficult to monitor or measure in CBE.

The assessment must be conducted concurrently with the progression of the training scenario to ensure objectivity; otherwise, the final evaluation at the conclusion of the training scenario may become subjective as the assessors may not have a complete recollection of all events. Experience of seafarers and their perceived realism of simulator training supress the effect of measurement.
3 RESULTS

3.1 RQ1: How differing assessment methods are operationalized in maritime simulator training?

To explain how the different assessment methods are operationalized in simulator training, the findings were categorized in three (03) segments:
- assessment goals,
- assessment tools, and
- assessment methods

3.1.1 Assessment goals

Apart from the differences of theoretical concepts, assessment in maritime simulator training has a wide variance of application in practice as revealed during the analysis of 18 selected studies. The goals of assessment in maritime simulators as identified based on their frequency of appearance in these studies are calculated. The analysis reveals that assessing navigational competence (33.3%) and situational awareness (28%) are the most focused assessment goals, followed by assessing human error (11%), non-technical skills (5.6%), lifeboat launching skills (5.6%), cognitive workload (5.6%), visual attention (5.6%) and others (5.6%). Bridge simulators (66.7%) are most used types of simulators in assessment contexts followed by engine plant simulator (16.7%), crane simulator (11.1%) and lifeboat simulators (5.6%) (see Figure 2).

3.1.2 Assessment tools

The results pointed out the predominance of questionnaire (32%) and observation techniques (28%) as assessment tools, where evaluators’ subjectivity and expertise played a crucial role in the overall judgement. Alternatively, heart rate variability and workload analysis from ECG (electrocardiogram) and EEG (electroencephalogram) signals, eye fixation duration and counts from eye tracking data, and algorithm-based analysis were sighted as quantitative measurement techniques (see Figure 3). However, irrespective of the tool or parameter used, the evaluators were involved either in the scale development or in determining what constituted good or bad performance, resulting in a relativist process of assessment [10]. This involvement nullified the potential objectivity of the seemingly realist quantitative techniques i.e., use of a criterion-based assessment.

3.1.3 Assessment methods

The analysis of the selected literature suggested that targeted behavioural observation and hands-on testing are the most used assessment methods depicting the prevalence of skill-based skill training in maritime simulator training. Other methods such as embedded measurement, structured situational interviews, secondary task performance and a few novel methods (i.e., EEG for measuring heartrate variability) are operationalized for evaluating skill-based learning outcomes. Measuring cognitive learning outcomes include self-report, power test, probed protocol analysis, speed test, recognition and recall methods of assessment. On the other hand, self-report measures are the solely used method for measuring affective learning outcomes in maritime simulator training contexts (see Figure 4).
3.2 RQ2: What are the advantages and limitations of operationalized assessment methods in maritime simulator training?

The review revealed context-specific advantages and limitations of differing assessment methods. For example, targeted behavioural observation is predominantly used for subjective evaluation; however, it facilitates objective evaluation if coupled with eye tracking data or other computer-based assessment measures. Similarly, hands-on testing facilitates instructors’ crucial input in different stages of learning which may be considered a disadvantage if the overall goal is to provide more objective assessment. Henceforth, a succinct overview of the pros and cons of every assessment approach, as implemented in differing maritime simulator training context, is presented:

3.2.1 Advantages

State-of-the-art tools such as eye-tracking, EEG exhibits novel promises in training evaluation facilitating objective and continuous observation of trainees [16], [33]. Computer-aided assessment tools (e.g., CAPA) demonstrate inter-rater reliability potentially reducing the involvement of instructors and subsequent bias in the assessment process [3] while other computer-based methods assure instructional support for the users [26], both aiming for objective evaluation of trainees. Other targeted behavioural observation and hands-on testing instances allows for continuous monitoring of trainees [33] as well as providing novel insights especially while using sensor fusion and biometric data [35] in measuring situational awareness (SA). On the other hand, the methodological characteristics of several assessment protocols (e.g., hands-on testing, probed protocol analysis) allows for more instructor involvement facilitating expert-in-the-loop and efficient subjective evaluation during training [15], [37], [38]. In addition, subjective methods (e.g., self-report) are found particularly useful to administer with a wider population of trainees especially during SA measurements [36].

3.2.2 Limitations

The identified limitations as described in respective studies utilizing various assessment methods can be categorized broadly in two classes namely hardware-based limitations and methodical limitations. The former encompasses difficulties arising from the resource intensive processes such as expensive eye-trackers as well as limitation of analysis software in dynamic assessment situations [32], [33]. On the other hand, methodical limitations encompass a range of issues such as insufficient assessment time [31] or being too time consuming [34], lack of variation in scenarios [12]; unintentional favouritism or bias [37], [38], familiarity effects [36], immeasurable behavioural constructs [26], [36], unclear evaluation criteria [26], ambiguity among instructors about assessment tools and procedures [27], subjectivity of these tools [3], [27] and related validity concerns [3], [35].

4 DISCUSSIONS

The focus of this study is to explore different assessment methods with regards to their objectives, suitability to the learning outcomes (i.e., cognitive, skill-based, and affective outcomes) and context-specific advantages and limitations. Based on the results of this study, we discuss the state-of-the-art maritime simulator training and assessment and propose a few future extents where the emerging assessment methods should focus on.

4.1 State-of-the-art maritime training and assessment

The concept of simulator training stems from the notion of competency-based training facilitating seafarers’ knowledge and skill acquisition required for their professional work [2], [43]. However, this review highlights the ubiquitous goal of simulator training being navigational competency training and situational awareness assessment in bridge simulators (see Figure 2) while other forms of competency training (e.g., engine room operators’ training, emergency procedural tasks, non-technical skill training etc.) are less prevalent in maritime institutes. Thus, understating specific competencies during simulator training ultimately defeats the purpose of all-round competency development for seafarers, while also contradicting the goal of authentic training in simulators, where the actual work environment may require seafarers to be competent in a diverse set of skills.

The results also suggest that a handful of assessment methods are frequently used while other prescribed methods such as the ones suggested by Kraiger et al. (1993) are underutilized. For example, targeted behavioural observation and hands-on testing are most prevalent while other types of assessment methods are barely utilized in maritime simulator training context. This could be due to the disproportionately large focus on measuring cognitive and skill-based learning outcomes in maritime simulator training while emphasizing less on affective learning outcomes. Nevertheless, from a human
factor’s perspective, affective learning outcomes are found to be correlated with real-world performance. Maritime research highlights the importance of affective components such as emotion [44], self-efficacy, motivation [23] and attitude in maritime training [45], [46].

In addition, instructors find themselves using traditional tools such as questionnaire and observation techniques prevalently in maritime simulator training despite the growing evidence related to the utility of physiological measures in performance assessment such as EEG [16], eye-tracking [32], [33] and functional Near-Infrared Spectroscopy (fNIRS) technology [47] in measuring cognitive resources such as workload, attention, stress etc. In addition, novel assessment protocols utilizing deep learning algorithms (e.g., artificial neural network) are envisaged as promising in categorizing trainee performance during maritime simulator training [48]. The instructors’ lack of interest in implementing new assessment methods could be attributed to either their unfamiliarity with such methods or concerns over their validity.

4.2 The future of assessment in maritime simulator training

This review has identified a significant gap in the assessment of maritime trainees’ affective learning outcomes (e.g., motivation, self-efficacy etc.), indicating a need for further research to explore innovative approaches to measuring such outcomes in the context of simulator training. This concept aligns with the conclusions drawn from other scholarly investigations as well [38], [49]. Therefore, future studies should prioritize investigating new and effective methods for assessing affective learning outcomes in this domain.

The goal of simulator training is to address the issues related to the lack of authentic learning and assessment contexts [50] while the assessment aims to bridge the gap between theory and practice [15], [22]. However, the lack of “workplace-relevance”, i.e., authentic assessment protocols is a well-known issue in maritime training context [6], [24]. Also, the effect of learning depends on the level of authenticity in simulator training [50]. Future studies should focus on establishing authentic assessment protocols satisfying the required criteria such as authentic training context [51], having real-world relevance and facilitating opportunity for collaboration [52] along with its seamless integration with the training activity [51].

The review highlights the subjective nature of traditional assessment tools such as questionnaires and observations, and the challenges associated with implementing them, including high costs (e.g., expensive eye-tracking equipment) and time-consuming processes (e.g., long briefing and debriefing sessions). Moreover, the instructors may lack experience or clarity on evaluation criteria [26], exacerbating these challenges. To address these issues, there is a growing trend towards using more objective and standardized methods for assessing maritime trainees in simulated environments [3], [16], [47]. Such methods are deemed beneficial in reducing subjectivity and bias while providing deeper insights into trainees’ performance. Research has also identified that students correlate fairness of assessment with their own engagement in the process [53] which is often ignored in the current assessment practices in maritime simulator training. Future studies should focus on the investigating novel learner-centred methods focusing more on instructor-trainee collaboration, validating emerging tools and methods through empirical research to increase instructors’ confidence in using them.

The discussions above enunciate that future assessment methods should focus on all learning outcomes, be authentic and integrated into realistic training contexts, and be practical in terms of cost and time efficiency. These methods should also be easy to understand and to administer by the instructors. Additionally, it is important to incorporate teacher-student collaboration in assessment to reduce bias while still retaining expert input from instructors (see Figure 5).

5 CONCLUSIONS

This review focuses on exploring how differing assessment methods are operationalized in current maritime simulator training practices. Also, the suitability of various assessment methods in differing contexts, taking into consideration the learning outcomes in terms of cognitive, skill-based, and affective competencies are investigated. Our findings demonstrate that while some assessment methods align well with these learning outcomes, there is a lack of methods to measure affective competencies. Additionally, there is an overemphasis on navigation training at the expense of other competencies, which could hinder the all-round competency development of seafarers. Furthermore, there are existing challenges in operationalizing various assessment methods. Based on our review, a detailed analysis of current assessment methods is presented to propose an envisaged best-practice for the future, considering their specific advantages and limitations and
identifying areas for improvement. This analysis will enable maritime simulator instructors to select appropriate assessment methods, design assessment episodes to capitalize on their advantages while avoiding potential drawbacks and adapt their training to meet the needs of their students. Overall, it is essential to prioritize outcome-based, authentic, practical, and collaborative assessment methods to enhance the effectiveness of maritime simulator training.

ACKNOWLEDGEMENT

The authors acknowledge the support Centre of Excellence in Maritime Simulator Training and Assessment (COAST) in Norway, funded by the Directorate for Higher Education and Competence (HK-dir). The 1st, 2nd, and 4th author appreciate the support of the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 823904 (Project: ENHANCing Human Performance in Complex Socio-Technical Systems, ENHANCE).

REFERENCES


