Technology, Knowledge and Learning
https://doi.org/10.1007/s10758-024-09735-4

ORIGINAL RESEARCH

Review on Competency Assessment Instrumentation in Computer-based Simulation

Simen Hjellvik[1] · Steven Mallam[1] · Marte Fannelø Giskeødegård[2] · Salman Nazir[1]

Accepted: 31 March 2024

Abstract
Computer-based simulation is utilised across various educational fields, employing diverse technologies to facilitate practical understanding of content and the acquisition of skills that can help close the gap between theory and practice. The possibility of providing scenarios that resemble on-the-job tasks, enables instructors to both train and assess the trainee's comprehension of the tasks at hand. The practices as well as the technologies for the assessment of simulation-based training vary across disciplines. Our motivation is to address quality procedures from a cross-discipline perspective. There seems to be a lack of scientific investigation that takes one step back from the specific application and investigates how assessment instruments can be developed to fit training outcomes regardless of the professional discipline. This scoping literature review on empirical studies aims to do so by exploring how competency is assessed with computer-based simulation. Objectives to achieve this are: (1) apply established training research theory to structure a decomposition of assessment instruments; to (2) review approaches to assessments factored over this structure; and (3) discuss quality procedures taken in the creation of the reported instruments and then propose an approach to assessment instrumentation that can be applied independent of discipline, with the range of current technology, and for any focal outcome competency. By reviewing a spectrum of fields, we capture reported assessment practices across a range of currently employed technologies. This literature review combines the methods of a scoping review with the qualities of a systematic literature review while keeping to conventional reporting guidelines. This allowed us to provide insight into current approaches and research designs that applied measurements in the range from automated assessment to observer rating of simulation-based training in professional work settings. This study found that all reviewed studies measured skill-based outcomes with some variation and that there is more theoretical and empirical work to be done to close the gap on quality instrumentation and its validity evidence. Our contribution to the field of training research is the operationalized component structure and the synthesised approach to instrumentation that could offer researchers and practitioners guidance and inspiration to develop and conduct quality assessments in competency development.

Extended author information available on the last page of the article

Published online: 03 May 2024

Keywords Assessment Instrument · Competency Assessment · Computer-based Simulation · Simulator Training · Simulator Assessment

1 Introduction

Simulations serve as a basis to provide activities that develop knowledge, skills, and attitudes in a systematic training programme in a variety of professional disciplines (Grossman et al., 2014). Finding appropriate means to assess the performance of the trainees becomes a critical matter, including what information this assessment should be based on and how it is linked to the stages in the trainee's learning process.

One of the key challenges in designing and implementing simulation-based training is determining the most effective approach for conveying domain knowledge. This involves selecting suitable instructional methods and technology. It also entails making suitable decisions on how to evaluate the effectiveness of the training and whether it is successful in conveying the necessary skills to the trainees. The required competence needed to make sound choices is not necessarily owned by one person alone, as it will involve different fields of expertise. A challenge with the design and delivery of simulation-based training is to combine the roles of (1) subject matter expertise, (2) training instruction, and (3) technical simulator and software expertise (Hjellvik & Mallam, 2021). In addition, the role of instruction is also concerned with performance appraisal, that is, assessing the training outcome. In the context of providing discipline-specific training, one person rarely holds all competencies in these roles alone (O'Donnell et al., 2015). As such, in the development of a training design, there is also a challenge to develop and apply appropriate instrumentation that can assess the outcome of training and, in turn, bridge the performance exhibited during training with subsequent professional conduct (Passmore & Velez, 2014, pp. 136–153).

In competence development, simulation-based training could be a tool in a cycle from (1) a qualification requirement, the gap from this to (2) an identified need, followed by (3) the initiated intervention to close the gap, and finally (4) the effective change on-the-job. It is essential to evaluate a learner's performance in a suitable manner in order to determine their ability to accomplish their current training task and their problem-solving skills for future situations. Historically, the assessment has been carried out using instructors' professional judgement from observation, which operationally has been criticised for falling short of actual performance (Spence & Baratta, 2014).

Competence can be viewed as the collective knowledge, skills, abilities, and attitudes required to successfully perform a certain task in alignment with predetermined requirements and objectives. (Miller, 1990). This definition demonstrates that certain aspects of competence are more challenging to quantify compared to others, as some may involve tacit and implicit components of doing a "good job". However, certain individuals are more explicit and expressive and may be easily identified and defined using specific criteria in an assessment instrument. In this respect, it might be useful to distinguish between competence and competency, where this article focuses on the latter. Competence relates to qualifications while highlighting a link between individuals' capability and their competencies (Evans & Kersh, 2014). While often being used interchangeably, the term *competency* can refer to the professional behaviour that is derived from learning and training knowledge, skills, and attitude components towards an advanced ability to perform a specific task in a

professional manner (Carraccio et al., 2002). Evans and Kersh (2014) remark on the term *competency* in a *"performance-related sense as an element of vocational competence,"* where skills are linked to performance through tasks *"subject to subsequent measurement of the intended consequences [or learning outcomes]."*

With a competence requirement as the educational quality standard, the outcome of training within this education could be assessed with instruments that compliment it as evidence of competency. In contemporary simulator training, it is essential to have the capability to integrate available characteristics into instruments for evaluating proficiency. Moreover, the quality of instrumentation becomes a sociotechnical concern due to the need for a customised assessment that takes into account the specific competency being tested, the type of simulator technology used, the characteristics of the trainee population, and the assessment conducted by the instructor.

1.1 Competency Assessment

The variation in training outcomes, trainee population, simulator technology, and training scenarios poses key challenges for creating suitable instruments to assess performance. That is, the technical capabilities of the different simulator designs can limit the opportunities to tailor assessments of training outcomes to the desired training exercises. To illustrate, some solutions are delivered without any supporting assessment functions, thus completely relying on observer rating, while others have pre-programmed tasks and assessments that cannot be modified. A central challenge is also to create a structured instrument for assessment that allows one to define, measure, and assess professional competency across and between trainee populations. It could also be transferable across condition variations. Just as professional skills can be acquired through both apprenticeship and simulator training, so too can the outcome competency of training be examined through simulation or in real-life practice.

Simulation-based training is applied to outcome-based education that must integrate a vocational element with academic qualifications. In medical education, this has evolved into a trainee- and outcome-oriented organisation where specific milestones or benchmark competencies are expected to meet the respective competence requirements (Frank et al., 2010). Similarly, in maritime professions, competence requirements are met through training on vocational competencies comprised of task-based skills (Manuel, 2017). Manuel (2017) views this integration of vocational and academic education as a *"new university paradigm of merging inquiry and task-focused, outcomes-based educational approaches, [where] learners should be optimally challenged to question the status quo, to develop critical skills that in the main are cognitive while at the same time meeting the demands of specific competences related to specific professional standards".*

Assessment of competency, as such, can be based on pre-defined threshold standards or milestones rather than comparing one trainee with another. In the traditional vocational apprenticeship, professional skills are learned firsthand on a time-basis by following a mentor in an on-the-job setting and subsequently being rated. Whereas in a competency-based practice, for example, in medical education, skills are first acquired through simulation-based training and then assessed against a competence requirement (Wagner et al., 2017), before the trainee is allowed to be involved in a specific level of patient care (Chetlen et al., 2015). This aims to ensure post-graduation expertise that holds a professional capacity to handle any variety of scenarios with mastery or perfection (Holmboe et al., 2010). There

are three types of tools that Chetlen et al. (2015) identifies to evaluate skills in medical simulation-based training: (1) knowledge tests before and after the training; (2) checklists for procedural skills; and (3) psychometric tools for interpersonal and communication skills.

1.2 Current Knowledge

According to our preliminary investigation on the current literature, there have been several reviews on assessment methodology and application of simulator technology in recent years. Seemingly, the field of medicine stands out as the predominant field of research. The healthcare domain in general applies simulators across the domain for professional education with favourable effects, although the review of Ryall et al. (2016) on the effectiveness of simulation-based training finds that more research is needed to validate the effectiveness of stand-alone assessment instruments. An example of this is the Objective Structured Assessment of Technical Skills (OSATS), which van Hove et al. (2010) in their review of technical surgical skills, find to be the most applied assessment instrument; however, they too find its application and evidence in the operating room to be limited. To illustrate how evidence of validity might be operationally defined in simulation-based assessment research, Cook et al. (2014) reviewed 217 healthcare studies and found construct evidence to be the main reported factor for the validity argument. Interestingly, the largest type of construct validity reported was *"how simulator scores varied according to a learner characteristic such as training status,"* while the second largest was expert-trainee discrimination. Also, separately measured concurrent or predictive variables were frequently reported. In their review of skill transfer from simulation-based training, Dawe et al. (2014) investigated the correlation between simulated performance and surgical performance. By reviewing randomised controlled trials (RCT), they found evidence that proficiency acquired through simulation-based training performed better than their patient-based training counterparts. The assessment criterion for these RCTs was global rating scales, such as the OSATS, meaning the criterion was based on observer rating without reporting any integration of digital parameters from the simulators. In sum, concurrent meta-research on simulation-based assessment offers both examples and critiques of task-specific assessment instruments used for training professional competencies. However, there seems to be a lack of reviews that takes one step back from the specific application and investigates how assessment instruments can be developed to fit training outcomes regardless of the professional discipline.

1.3 Research Objectives

Our preliminary overview of competency assessment with computer-based simulation shows there are a variety of practices to ensure quality. When applied to specific training outcomes beyond the scope of their design, standalone instruments like global rating scales could offer limited flexibility for new and broader applications. They also rely on observer ratings by design, which systematically limit the use of objective simulator data when available. For the validation of instruments, construct validation may not be sufficient alone, as training for a test is something different than training for real-world scenarios. Expert-trainee discrimination might be a useful quality procedure for an assessment instrument; however, a concurrent measurement might be preferable to add, as experts likely did not develop into agents of the profession through the same technologies as the current trainees.

To review reports from studies up to the present day raises interest in this regard, especially considering that all the reviews were limited to their distinct disciplines.

This review is positioned to expand the reported practices from different disciplines to a generic application. We pose the research question: "*How is competency assessed with computer-based simulation?*". To address this research question, a review of current applications of assessment practices follows. The motivation for this study is to investigate quality actions in instrumentation currently employed in research on professional education and training. With a variety of fields incorporating computer-based simulation, cross-disciplinary scoping is needed to find commonalities and salient qualities that can be redeployed to benefit future practice. Thus, our aim is to investigate how competence requirements in various disciplines are measured during and after the training of professional competency.

In answering the research question, we adopt the following structure with objectives over three phases that will (1) provide a theoretical foundation to operationalize components of assessment instruments to organise the review output. Then, (2) perform a systematic literature review and describe the characteristics of the procedures found. And finally, (3) propose an approach to assessment instrumentation based on quality characteristics found in the review.

2 Methods

The nature of the research question calls for a theoretical exercise to make propositions for a subsequent empirical follow-up. First, training research was investigated to prepare for the review process and provide a theoretical foundation. Section 2.1 addresses the first objective and translates theory on assessment instruments into operational components and the dimensions of the review. This was necessary for organising the data extraction part of the review process and the subsequent synthesis of data.

Then, to explore the research question, "*How is competency assessed with computer-based simulation?*" The literature review was initiated through three phases; (1) search, (2) selection, and (3) extraction, leading to the evaluation and interpretation of the data in the context of the selected dimensions (Randolph, 2009).

This study employed a scoping review, which is characterised by the pursuit of defining boundaries around a particular topic to inform future primary research (Sutton et al., 2019). Scoping the available research literature can be helpful to identify the nature and extent of prior and current research evidence; however, the elements of quality assessment typically found in a systematic review were included (Grant & Booth, 2009). Best practices for a systematic review were followed to ensure transparency and reproducibility (Snyder, 2019).

The resulting data is presented in the next chapter and synthesised to meet our third objective in the discussion chapter.

2.1 Assessment Instrument Organization

In this section, we apply established training research theory to structure a decomposition of assessment instruments. This provides the dimensions of (1) outcome competency and (2) component level for the assessment instruments to be reviewed.

In training technical skill-based outcomes, an appropriate evaluation tool has been the observation of on-the-job performance and behaviour. Kraiger et al. (1993) built on this tradition and classified the *learning outcomes* of training into the groups of (1) cognitive, (2) skill based, and (3) affective outcomes. Cognitive outcomes comprise knowledge-based learning *categories*, from the initial stage of acquiring knowledge that can be verbalised, to organising knowledge and strategizing the application of it. As training is a process of learning, *learning constructs* subsuming the *categories* are interlinked as different measurements focus on the same intervention. For example, proceduralization is considered a skill-based construct and process that enables the building and reproduction of trained knowledge, application, and behaviours. This is like the cognitive constructs of organising mental models and strategizing metacognitive skills of self-regulation. The purpose of the framework is to guide training evaluation towards appropriate measurement of the focal constructs and ensure that the selected *learning constructs* correspond appropriately to the *learning outcome*. Affectively based outcomes concern attitudinal and motivational learning constructs that have been connected to performance. For example, safety-related competencies might have training objectives that focus on the strength and direction of attitudes.

In addition, the concept of (4) non-technical skills must be explicitly included to supplement the skill-based outcome class. The Kraiger et al. (1993) framework did not emphasise on non-technical skills as an independent class of *learning outcomes*, but this becomes an important facet of skill-based outcomes whenever training for an *outcome competency* that involves more than one individual operating a technical system. Non-technical skills are, by definition, *"the cognitive, social, and personal resource skills that complement technical skills…"* (Flin & O'Connor, 2017, pp. 1–16). To illustrate, situational awareness, decision-making, communication, teamwork, leadership, stress management, and fatigue resilience. In review, non-technical skills are not a mutually exclusive category within the framework. For instance, leadership, stress management, and fatigue resilience coincide with the *learning constructs* of self-insight and metacognitive skills under the cognitive strategies *category*, which are classified as cognitive *learning outcomes* with a measurement focus on self-awareness and self-regulation. Further, non-technical skills such as situational awareness and mental workload coincide with the *learning constructs* of automatic processing and tuning under the automaticity *category*, which are classified as skill-based learning outcomes with a measurement focus on attentional requirements and available cognitive resources. To operationalize a component structure for assessment in simulator training, Fig. 1 is proposed.

Figure 1 shows a structure that recognises outcome competency as the main goal and purpose of the specific learning process. It has several stages that are paired with the key performance indicators (KPIs) that are needed to describe the outcome. For example, declarative verbal knowledge, simulator performance, and resource management. KPIs relate back to the main categories of learning outcomes, whether cognitive, technical skill, non-technical skill, or affective dispositions. The KPIs cannot themselves be directly measured, but they can be accumulated by several subsuming facets of performance indicators (PI) corresponding to learning constructs. As such, PIs can be assumed or verified with evidence to validate the construct. In other words, PIs can be quantified or dichotomous learning constructs in different contexts to indicate or validate the presence of a KPI. To illustrate, knowledge might be examined before training, immediately after, or at a deferred point in time. Task performance in numerous repeated or different simulations might serve as a KPI for simu-

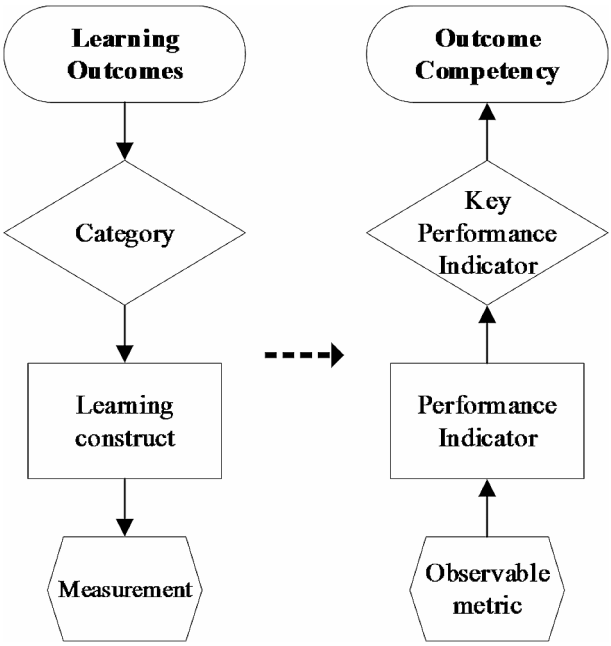

Fig. 1 Components of competency assessment from concept to evidence

lator performance. The foundational level of the Fig. 1 model concerns the quantification of observable evidence to describe the PIs under investigation. By instruments feasible to the targeted PIs, the observable evidence of objective metrics and parameters is the basis for competency assessment. If the assessment instrument is well organised, it constitutes credibility for the outcome competency of the training process and, by extension, the competence requirement that initiated the training. In summary, the component structure meets our first objective, where (1) outcome competency and (2) component level are used as the main dimensions in the following review process.

2.2 Search Strategy

The literature review question was formulated to describe feasible approaches to simulator assessment based on reported scholarly activities and provide an overview of how assessment could be structured for specific competencies assessed in both a virtual and on-the-job environment. Thus, a scoping review (Sutton et al., 2019) approach was selected with PRISMA process documentation (Moher et al., 2009), to collect the data to be categorised and address the literature review question. A search strategy based on search terms was developed from the PISCO framework (Schardt et al., 2007) to be operationalized with the individual electronic database search. The search engines SCOPUS, ERIC, World of Science (WoS), and IEEE, were selected to cover a broad spectrum of scientific fields. As these databases apply different functionalities to their search engines, index search terms corresponding to Table 1 were identified and translated into the respective index words of the search engines.

The individual search stings were formulated to be as equivalent as possible with consultancy from a senior librarian at a post-secondary institution. This resulted in prioritising the

Table 1 PISCO search terms

Population	Intervention	Setting AND/OR Comparison	Outcome
Adults Professionals Students Trainees	Education Knowledge Acquisition Learning Skill Acquisition Training	Education Professional Education Professional Training Simulation Simulator Training	Ability Assessment Competence Learning Performance Proficiency Skills

Table 2 Individual database search string

Database	Search string	Result
SCOPUS	(KEY (simulator*)) AND (KEY (assessment OR evaluation)) AND (KEY (training)) AND (LIMIT-TO (EXACTKEYWORD, “Simulators”) OR LIMIT-TO (EXACTKEYWORD, “Simulator”))	693
ERIC	DE “Simulation” AND DE “Evaluation”	257
WoS	TS=(“simulator* training”) AND TS=(assessment OR evaluation)	279
IEEE	((((Index Terms: simulator))) AND ((Index Terms: training))) AND ((Index Terms: assessment OR Index Terms: evaluation))	244

setting and outcome over intervention and population, as seen in the final search strings in Table 2. Aligned with our strategy to perform an inclusive and simple search, the two latter terms were not applied. Comparatively, this resulted in a larger body of potential articles ($n=1473$) to be assessed and selected. Ultimately, whether an article is discovered by the search strings or not can be biased by the keyword indexing made by the authors and the publishing journal. The search process was conducted in February 2022 and updated in March 2022.

2.3 Selection Criteria

Eligibility criteria were defined from the search terms (see Table 3) to govern the selection process. Language was restricted to English, published works of any date leading up to the conclusion of the search process in March 2022. Criteria one frames the population in a range equal to the able-bodied age of a generic profession, including driver licence proficiency training regardless of age. The authors decided that studies with simulators as a lower-school learning aid or with populations using simulators for health care purposes fell beyond the scope of training and education for professional trades and skills. Criteria two excludes records from before computer-based simulation started to appear in scientific articles and segregates the design and delivery of simulation from similar tools. Criteria 3 and 4 balance the selection based on the reported application of observational assessment tools with objective measurements, as is aligned with the purpose of this review. However, the use of only subjective instruments without the support of structured observational tools or non-observational tools was excluded. This is because the assessment of computer-based simulation should itself be at least partially computer-based or compatible with objective data to meet our objectives.

Table 3 Eligibility criteria

	Criteria 1	Criteria 2	Criteria 3	Criteria 4
Inclusion	Population: Adult students or professionals at UNESCO ISCED level 5 or above. Driver license students. Type: Peer reviewed journal articles and conference proceedings.	Intervention: Experiment or training with computer-based simulation capable of providing simulator metrics.	Setting / Comparison: Assessment or evaluation of performance by subjective, objective or hybrid approaches. Methodology and operationalization described. Adaption of generic methods.	Outcome: Performance assessment of single or repeated training exercises. Competence assessment of summative training/ learning. Performance indicators from programmed or observed metrics used for quantifiable assessment.
Exclusion	High School level or lower, elderly, disabled or patients.	Does not use simulator metrics. E-learning, technology-free or non-computer-based simulation.	Appropriation of task- or discipline-specific methods without any description of originality.	Subjective assessor rating or self-assessment without the combination of objective measures.

The preferred reporting items for systematic reviews and meta-analyses protocol (Moher et al., 2009) was adhered to when documenting the selection process (see Table 4). The process was managed with EndNote X9, and the included records were exported to Excel for further data extraction. The selection process stated with excluding duplicate records from the 1473 large initial pool. Record types other than journal articles and conference proceedings, e.g., books, book sections, serials, and theses, was excluded. Then, 1234 retaining abstracts were reviewed according to the exclusion criteria, where 684 records were removed. The search for full text versions rendered fifty-six records which proved unobtainable by any available effort, i.e., through university access, national universities collaboration access, and through direct request to the authors. An independent search in the full text pool disclosed some review articles and meta-analyses, with the later crite-

Table 4 Abbreviated PRISMA report card

Operation		n retained	n excluded
Identification	Database search merged	1473	
	Duplicates removed		71
	Excluded on TYPE		168
Screening	Excluded on abstract		684
	Full text search	494	56
	Review articles		56
Eligibility	Criteria Assessment	438	
	Excluded on criteria		403
Included	Included for review	35	

ria assessment the total number of such types grew to fifty-six. 438 journal articles and conference proceedings were assessed for the inclusion and exclusion factors, rendering thirty-five records to be included in the review. Examples of discarded records are studies in laparoscopic surgery training with a low technology or non-computer-based simulator where the performance is evaluated by a single expert's assessment of a video recording of the training, or computer-based simulation where the tasks and assessment metrics scoring system are default-programmed by the simulator developer and not described or justified in useful detail. The eligibility criteria assessment was performed by the first author alone with decisive support from the co-authors on hesitant matters.

2.4 Data Extraction and Quality Evaluation

The included thirty-five records were reviewed in depth, confirming the criteria and factors and highlighting the data to be extracted. This qualitative data was collected in a spreadsheet synthesis matrix in the factors (1) outcome focus and training structure, (2) assessment methodology, PIs, and measurements, (3) simulator equipment and design, and (4) research methodology, sample, and statistics. At the component level (see Sect. 3.2), we assigned each entry within the matrix to a dimension and assigned a subjective evaluation based on its level of approximation to the objective. We intended this rating to facilitate the synthesis of information for further discussion. Also, descriptive data and references to be followed up by a saturation search were collected. After all records were reviewed and their data extracted, the saturation search references was investigated. No additional inclusion in the review body was made based on these references, as they mainly elaborated on the methodology and instruments used. Four records were conference proceedings, and thirty-one were journal articles comprising different fields:

- 14 medical training.
- 6 maritime education and training.
- 5 automotive training.
- 3 military combat or military aviation training.
- 2 process plant operation training.
- 1 dental surgery training.
- 1 heavy machinery operation training.
- 1 railway locomotion training.
- 1 ambulance driving.
- 1 athletic training.

Table 5 Summary of review sources

Database	Search	Included	Index	Precision
SCOPUS	693	22	0.63	3,1%
ERIC	257	0	0	0,0%
WoS	279	9	0.26	3,2%
IEEE	244	4	0.11	1,6%
Σ	1473	35	1	2,4%

Table 6 Learning outcome classes of the reviewed studies

Cognitive $n=5$	Prohn and Herbig (2020), Sportillo et al. (2019), Sullman et al. (2015), Taylor and Barnett (2010), Taylor and Barnett (2013)
Skill-Based $n=35$	Bajka et al. (2010), Boyle et al. (2011), Bratko et al. (2020), Brunckhorst et al. (2015), Bube et al. (2019), Chang et al. (2016), Chowriappa et al. (2013), Colombo and Golzio (2016), de Winter et al. (2009), Duarte et al. (2013), Ebnali et al. (2019), Ernstsen and Nazir (2020), Hjelmervik et al. (2018), Iqbal and Srinivasan (2018), Konge et al. (2013), Li et al. (2020), Liu et al. (2020), Loukas et al. (2011), Mackel et al. (2007), Madsen et al. (2014), Nisizaki et al. (2017), Okazaki and Ohya (2012), Pagnussat et al. (2020), Poursartip et al. (2018), Prohn and Herbig (2020), Rauter et al. (2013), Rhienmora et al. (2011), Rosenthal et al. (2015), Scavone et al. (2006), Sportillo et al. (2019), Sullman et al. (2015), Taylor and Barnett (2010), Taylor and Barnett (2013), Verstappen et al. (2022), Ojados Gonzalez et al. (2017)
Affective $n=4$	Ebnali et al. (2019), Ojados Gonzalez et al. (2017), Prohn and Herbig (2020), Taylor and Barnett (2013)

As shown in Table 5, the ERIC database failed to produce any contribution to the review. Precision is the fraction of included records in relation to the database search, and index is the fraction of each database's contribution to the included pool. The majority of included records originated from the SCOPUS database, while the most accurate capture came from the Web of Science database search. The largest proportion of included records came from SCOPUS. While considering the criteria, the WoS and SCOPUS search strings were equally precise. Accumulated, all operationalizations of search strings yielded a 2.4% effectiveness.

3 Results

3.1 Outcome Competency Dimension

For an overview of the measurement focus found in the reviewed records Table 6 illustrate dissemination of the outcome competency dimension to the learning outcomes (1) cognitive, (2) skill-based, and (3) affective. Whereas all report some element of skill-based measurement, three of the studies incorporated measurements from all classes.

3.2 Component Level Dimension

In view of the component-level dimension of Fig. 1, the reviewed studies were factored as three types with an accretive level of complexity for the assessment schemes applied. Namely, observable metric comparisons ($n=17$), performance indicator assessments ($n=14$), and composite methods ($n=4$). These three categories are different in the way they have incorporated the structure levels into their assessments. Assorted designs of simulator technology were reported, including off-the-shelf commercial solutions and purposely built solutions; however, no cloud-based simulator was found.

3.2.1 Observable Metrics Comparisons

The first type of assessment comprises studies with a simplistic approach, considering the data collected for the assessment. The characteristics of these studies are summarised in Table 7 below. In brief, these compare a few or single metrics in (1) group comparison, (2) within-group comparison, or regressing on a dependent variable to identify predictor metrics. Common for these is that no aggregation of the collected metrics was reported; that is, the data was used as is. No systematic organisation of the metrics was applied. As such, the simulator-generated, measured, or observed metrics alone represent the focal PIs, KPIs, or competency outcomes under investigation. Quality characteristics and the relevance of these studies are discussed in Sect. 4.2.1.

3.2.2 Performance Indicator Assessments

The second type of assessment comprises studies with a structured approach. These are studies that collect multiple simulator-generated, measured, or observed metrics that were aggregated into PIs and KPIs through a scaled or dichotomous structure. For example, the use of dichotomous checklists to create an average KPI score, procedures with weighted scoring structures, or global rating scales that incorporate different KPI categories, PI constructs, and weighted metrics. Some studies also applied normalisation, i.e., rescaling metrics into variables with the same range, and some applied standardisation procedures, i.e., rescaling metrics by indexing the mean and standard deviation into a new variable. Some applied methods for analysing repeated measures, and some analysed group comparisons. The characteristics of these studies are summarised in Table 8 below. Quality characteristics and the relevance of these are discussed in Sect. 4.2.2.

3.2.3 Composite Methods

The third type of assessment is distinct by applying multiple or complex methods to assess a skill-based outcome. The only clear commonality across the four studies in this type of assessment was that all employed a between-group design. Technology also differed, as Ernstsen and Nazir (2020) used maritime full-mission bridge simulators, Rauter et al. (2013) used a full-mission rowing simulator, Rhienmora et al. (2011) used a dental 3D desktop VR simulator, and Mackel et al. (2007) used a sensorized female mannequin. The characteristics of these studies are summarised in Table 9 below. Quality characteristics and the relevance of these are discussed in Sect. 4.2.3.

Table 7 Characteristics of observable metrics comparison type studies

Study ($n=17$)	Simulator concept	Study design	Intervention	Assessment metrics	Outcomes
Bratko et al. (2020)	Wearable sensors	Testing of conceptual assessment methodology with undisclosed sample size.	Direct force-on-force action training.	Number of shots, number of hits, received injuries, and effectiveness of fire activity	Skill-based outcome by individual and team offensive or defensive performance.
Colombo and Golzio (2016)	3D desktop VR process plant simulator.	Between group with students training either through static graphic presentation or dynamic 3D presentations with passive 3D glasses ($n=24$).	Field operator simulation.	Hints, message repetition, leakage identification, valve identification, fire reporting, pool diameter, flame height, and total time.	Skill-based outcome by operator performance.
Ebnali et al. (2019)	Fixed-base platform with 120° projection screen.	Within- and between group with students given simulator training, video training, or no training control ($n=54$).	Autonomous vehicle driver training course.	Takeover Time, speed, speed variance, standard deviation of lateral position, takeover decision accuracy, trust, and acceptance.	Skill-based outcome by taking control over vehicle. Affective outcomes by attitudes towards autonomous vehicles.
Hjelmervik et al. (2018)	Full mission bridge simulator	Between group trained either with heterogeneous or homogeneous ocean currents ($n=17$).	Navigational and ship handling simulation.	Cross track deviation from route path.	Skill-based outcome by ship handling performance.
Iqbal and Srinivasan (2018)	2D desktop process plant simulator.	Between group with two proficiency levels of students ($n=128$).	Simulated ethanol production plant.	Time margin to failure, available time before shutdown, and response time to restore operation.	Skill-based outcome by reliability of control room operators.
Li et al. (2020)	Full Mission crane simulator	Between-group with crane operators ($n=12$).	Oil installation crane operation.	Saliency similarity of eye-tracked heat map comparison.	Skill-based outcome by attentional requirements.
Loukas et al. (2011)	3D desktop VR laparoscopic simulator with haptic feedback.	Between-group with resident- and expert surgeons, and within-group learning curve comparisons ($n=22$).	Repeated laparoscopic tasks and pre-/post-training tasks.	Errors in dexterity, safety, and technical skill from default-programmed simulator metrics.	Skill-based outcome by error rates of psychomotor skills.
Nisizaki et al. (2017)	Full mission bridge simulator	Between-subject ($n=7$).	Congested passage navigation.	Subjective performance checklist, situational awareness, mental workload.	Skill-based outcome by attentional requirements.
Ojados Gonzalez et al. (2017)	Motion platform and head-mounted display.	Between-group of students with "safety training courses," farmers with "experience in driving tractors," and students "without experience in driving tractors" ($n=127$).	Deployment of safety gear during tractor driving simulation.	Driving time; time stopped on route; number of times safety gear deployed, and route plan pointing to the places of the errors of item (3)	Skill-based outcomes by safety behaviour. Affective outcomes by perception of risk and safety.

Table 7 (continued)

Study ($n=17$)	Simulator concept	Study design	Intervention	Assessment metrics	Outcomes
Okazaki and Ohya (2012)	Full mission bridge simulator	Between-subject ($n=4$).	Congested passage navigation.	Situational awareness	Skill-based outcome by attentional requirements.
Pagnussat et al. (2020)	3D desktop VR forestry harvester simulator	Withing-group with random inexperienced participants ($n=12$).	Log harvesting simulator training course.	Measures of run time, fall direction, and cutting height	Skill-based outcomes by bimanual motor skills.
Prohn and Herbig (2020)	Unspecified driving simulators. One stationary and one mobile.	Within- and between group longitudinal with two intervention groups and one waiting control group ($n=183$).	Ambulance driver course.	Measurements at the levels of reactions to training, learning, and results of training.	Cognitive outcomes in terms of knowledge. Skill-based outcomes by driver performance. Affective outcomes by change in attitudes and behaviour.
Sportillo et al. (2019)	Fixed base VR driver simulator and augmented reality enhanced live driving condition.	Between-group (VR/AR/live) with hired consumer test participants ($n=60$).	Autonomous vehicle test drive.	Pre- and post-knowledge test. Reaction time to take over vehicle control from automation.	Cognitive outcomes by knowledge test. Skill-based outcomes by speed of performance.
Sullman et al. (2015)	Full mission bus body based on an electro-pneumatic motion platform with four axes of movement.	Between group professionals in either a treatment group given intervention, or a control group given a first aid course ($n=47$).	Eco-driving course.	Fuel consumption, distance driven, time taken to complete the drive, eco-driving knowledge, mental workload, and simulator face validity.	Cognitive outcome by knowledge test on subject. Skill-based outcome by eco-driving transfer measures.
Taylor and Barnett (2010)	3D desktop VR simulator, and a wearable simulator with head-mounted display and assault rifle.	Between-group with students ($n=98$).	Tactical movement, selecting fighting positions, and use of frag grenades.	Training retention test, mental workload, and motivation.	Cognitive outcomes by declarative knowledge retention. Skill-based outcomes by available cognitive resources.
Taylor and Barnett (2013)	3D desktop VR simulator, and a wearable simulator with head-mounted display and assault rifle. Live analogue simulation group in study 3.	Study 1: 8 evaluators assessing the simulators. Study 2: Between-group with students ($n=98$). Study 3: Between-group (desktop, VR, live) and within-group (multiple scenarios) with non-military participants ($n=62$).	Basic combat movement procedures training and hostage rescue scenario.	Knowledge test (study 2). Correct/incorrect performance of task steps in scenarios and live test (study 3).	Cognitive by knowledge retention. Skill-based by transfer of procedural skills. Affective by perceived performance.

Table 7 (continued)

Study ($n=17$)	Simulator concept	Study design	Intervention	Assessment metrics	Outcomes
Verstappen et al. (2022)	Full mission locomotive simulator	3 by 3 within-group with train drivers ($n=28$).	Locomotive operator safety training at different complexity levels.	Mental workload, attention allocation by eye-tracking and safety performance by simulator metrics.	Skill-based by safety performance in operation.

4 Discussion

In general, performance on discipline-specific tasks can be indicated by an assessment scheme of performance indicators that compose the task goal. Consequently, it may be assumed that the validity of measured performance in any simulation-based training is contingent on the quality of the assessment instrument. In view of the issue with measurements being transferable across conditions, this might induce a necessity to consider other factors of measurement than what can be derived from simulator parameters alone. The extracted data was synthesised chronologically by the dimensions used to organize the results. We performed the synthesis based on (1) common characteristics, such as experimental design and units of measurement, and (2) particular characteristics, such as the method and result of application. Then, (3) characteristics and qualities were discussed. Finally, we take one step back to address and discuss the objective of proposing an approach to assessment instrumentation.

4.1 Outcome Competency

It is interesting to note that there was variation across outcome foci. However, each study evaluated performance to some extent, with most focusing only on outcomes related to skills. Four studies reported affective effects, while five studies provided cognitive outcomes, according to Table 6. Descriptive knowledge was the main emphasis of those assessing cognitive outcomes, whether it was during training, immediately after training, or at some later time after training ended. These studies used recognition or recall tests to measure knowledge, as is conventional. The distinctions between skill-based and cognitive outcomes can seem subtle since they only represent different viewpoints. For instance, Ebnali et al. (2019) measured decision accuracy and labelled it as a higher-order cognitive skill. In view of the Kraiger et al. (1993) frame, decision accuracy, or reduction of decision error, comprises both the skill-based constructs of *proceduralization* and *composition*. Through the process of *proceduralization*, learners build distinct reactions and behaviours based on learned knowledge, thereby decreasing errors. Later, through *composition*, learners integrate these procedures into a complex system to perform in a fluid manner. Further, decision-making is, according to Flin and O'Connor (2017) a non-technical skill. Although all thirty-five studies reported measures of skill-based outcomes, only a few focused on the typical non-technical skills. Namely, available cognitive resources (Colombo & Golzio, 2016; Liu et al., 2020; Nisizaki et al., 2017; Sullman et al., 2015; Taylor & Barnett, 2010, 2013; Verstappen et al., 2022), attentional requirements (Brunckhorst et al., 2015; Li et al.,

Table 8 Characteristics of performance indicator assessments type studies

Study ($n=14$)	Simulator concept	Study design	Intervention	Assessment metrics	Outcomes
Bajka et al. (2010)	3D desktop VR hysteroscopic simulator.	Between- and within-group with novices and experts ($n=36$).	Five repetitions of two different exercises on exploration and diagnosis.	Fifteen simulator metrics within the indicators: visualization, ergonomics, safety, and fluid handling.	Skill-based outcomes by diagnostic hysteroscopy skills.
Boyle et al. (2011)	3D desktop VR vascular intervention simulator.	Between- and within-group with nonexpert feedback, expert feedback, and control ($n=18$).	Six renal artery angioplasty/stenting procedures.	Time, volume of contrast, balloon size, balloon placement, stent deployment, and handling error. Video-based assessment.	Skill-based outcomes by complex endovascular skills.
Bruncknorst et al. (2015)	3D desktop VR ureteroscopy simulator.	Between-groups with medical students ($n=32$).	Ureteroscopy simulator training.	Time, time to ureteral orifice catheterisation, stone withdrawal, and stent insertion. OSATS. NOTSS.	Skill-based outcomes by ureteroscopy surgery skills.
Bube et al. (2019)	3D desktop VR urological simulator.	Between- and within-group with novices, intermediates, and experts ($n=49$).	Procedure in transurethral resection of bladder tumours.	Time use, percentage of bladder inspection, percentage of tumour resection.	Skill-based outcomes by procedural skills.
Chang et al. (2016)	3D desktop VR arthroscopic simulator.	Between-group with students, residents, and staff ($n=19$).	Diagnostic arthroscopy procedure.	Default-programmed simulator motion metrics of camera, probe, grasper distance, roughness in millimetres, and force in Newtons. Live observer evaluation.	Skill-based outcomes by arthroscopic surgery skills.
Chowriappa et al. (2013)	3D desktop VR robotic surgery simulator.	Between-group with non-robotic surgeons and on expert-robotic surgeons ($n=27$).	Five different robotic surgery training exercises.	Ten simulator metrics composited into performance indicators: bimanual dexterity, economy, critical errors, safety in operative field, and task time.	Skill-based outcomes by robotic surgery skills.
de Winter et al. (2009)	180° projector full mission car simulator with fixed base and vibration feedback.	Within-group of learner drivers ($n=804$).	Learner driver training course.	Simulator metric aggregated for scoring indicators: speed, speed- and safety violations, and steering error.	Skill-based outcomes by driver proficiencies.
Duarte et al. (2013)	3D desktop VR laparoscopic simulator.	Within-group with novices and no control ($n=11$).	Basic laparoscopic training tasks.	Error in camera handling, peg, and transfer, clipping, and cutting.	Skill-based outcomes by psychomotor skills.
Konge et al. (2013)	3D desktop VR bronchoscopy simulator.	Between-group with 3 groups of respiratory physicians ($n=22$).	Repetition of two standardized procedures.	Successful samples performed and procedure time aggregated to a score.	Skill-based outcomes by procedural skills.

Table 8 (continued)

Study ($n=14$)	Simulator concept	Study design	Intervention	Assessment metrics	Outcomes
Liu et al. (2020)	Full mission bridge simulator.	Within-subject of navigation officers ($n=4$).	60 min pilotage exercise.	Electroencephalogram (EEG) data for stress and mental workload detection.	Skill-based outcomes by psychophysiological resources during performance.
Madsen et al. (2014)	3D desktop VR gynaecological transvaginal ultrasound simulator.	Between- and within-group with novices and experts ($n=28$).	Repetition of seven training modules.	Fifty simulator metrics from seven training modules.	Skill-based outcomes by examination skills.
Poursartip et al. (2018)	Sensorized instrument laparoscopic simulator.	Between-group with novices and experts ($n=30$).	Suturing and knot-tying tasks.	Potential- and kinetic-energy applied on the surgical instruments.	Skill-based outcomes by psychomotor skills.
Rosenthal et al. (2015)	3D desktop VR laparoscopic simulator with haptic feedback.	Between-group. Study 1: children, residents, and surgeons ($n=43$). Study 2: free- or structured training novices, and experts ($n=69$).	Basic training tasks and laparoscopic cholecystectomy procedure.	Default-programmed metrics composited into performance indicators: accuracy, time, and economy of movement.	Skill-based outcomes by laparoscopic surgery skills.
Scavone et al. (2006)	Human patient computerized mannequin.	Within- and between-group with medical students ($n=16$).	Anaesthetic procedure for emergency caesarean delivery.	Multiple metrics and checkpoints subsuming the indicators: preoperative assessment, patient care, equipment check, intubation, and interoperative management.	Skill-based outcomes by procedural skills.

2020; Nisizaki et al., 2017; Okazaki & Ohya, 2012) and decision making (Brunckhorst et al., 2015; Colombo & Golzio, 2016; Ebnali et al., 2019). There was little overlap in the four studies that reported assessments of affective outcomes. Namely, they focused on constructs like motivation, safety of performance, perception of risk, and trust in automation. In sum, although the majority focused on facets of performance related to the context of the individual study, a wide range of outcomes were identified. This demonstrates the flexibility of computer-based simulation as a tool for competency development.

4.2 Assessment Methodology

4.2.1 Observable Metric Comparisons

When taken at face value, assessments based on individual raw or modified data points may have content validity; however, they might not accurately capture the phenomenon of the outcome competency. That is to say, the empirical item measured by itself may not fully capture the veracity of the outcome competency. As shown in Table 7, this type of assessment was conducted with seventeen studies from the review sample. The most typical method for obtaining these measures is by self-reporting, observation, or simulator systems,

Table 9 Characteristics of composite assessments type studies

Study ($n=4$)	Simulator concept	Study design	Intervention	Assessment metrics	Outcomes
Ernstsen and Nazir (2020)	Full mission bridge simulator.	Between-group of navigational officers ($n=16$).	Assessment of a pre-recorded pilotage scenario.	Two-factored outcome with multiple KPIs, PIs, and twenty observed metrics.	Skill-based outcomes by pilotage performance.
Mackel et al. (2007)	Sensorized partial mannequin.	Between-group with novice and expert gynaecologists ($n=30$).	Gynaecological examination.	five-dimensional binary vector associated with five pressure sensors.	Skill-based outcomes by examination skills.
Rauter et al. (2013)	CAVE type full mission rowing simulator.	Between-group of recreational rowers ($n=8$).	Training sessions with licensed rowing instructor.	Quantitative biomechanical performance metrics video evaluation	Skill-based outcome by rowing performance.
Rhienmora et al. (2011)	3D desktop VR dental simulator.	Between-group with student and experienced dentist ($n=10$).	13-stage crown preparation procedure.	Distinct instrument force and movement pattern.	Skill-based outcomes by procedural and psychomotor skills.

and then directly comparing them across demographic or intervention groups. Some studies reported a within-group design with learning curve analysis, while most studies applied a between-group design with differing expertise levels. The simulator equipment used was a head-mounted display ($n=5$) with virtual reality or augmented reality, full mission designs ($n=6$), 3D desktop VR ($n=5$), and one with a 2D desktop design.

Within-group design was reported as repeated measures, mixed methods, or with a series of scenarios in six of the studies (Ebnali et al., 2019; Loukas et al., 2011; Pagnussat et al., 2020; Prohn & Herbig, 2020; Taylor & Barnett, 2013; Verstappen et al., 2022). The power of a repeated measures design lies in its ability to potentially identify the point at which a learning curve transitions to a plateau. This would determine the point at which more training does not provide any additional value to the dependent variable at the chosen level of significance. Loukas et al. (2011) employed Freidman's analysis of variance (ANOVA) to identify the points at which the learning curve plateaued for all simulator metrics with

a pairwise comparison. Results demonstrated the point at which proficiency reached its maximum level when the laparoscopic simulator was used with a variety of repetitions, as indicated by the collected metrics. In a mixed method 3×4 ANOVA, Taylor and Barnett (2013) also used pairwise comparison to find differences in the grouping factor and the scenario factor. During that particular analysis, the examination of time duration successfully revealed the distinctions among the training conditions: desktop, wearable, and live. Additionally, it demonstrated the learning curves for the desktop and wearable conditions across four scenarios, while indicating a consistent curve for the live condition.

The convenience of assessments based on comparing metrics is that raw data can be analysed directly. The between-group designs mostly involve comparing two or more levels of expertise. These assume that the measures collected accurately reflect the level of competency, as shown in Fig. 2. This can only be true in cases where (1) the expert group truly is an expert in the outcome competency, (2) the observable metric captures and covaries with the outcome competency, and (3) the outcome competency transcends to the real-life condition where expertise is defined. Only five studies (Loukas et al., 2011; Ojados Gonzalez et al., 2017; Prohn & Herbig, 2020; Sullman et al., 2015; Verstappen et al., 2022) used experts or professionals that could be defined as experts of the outcome competency. Defined expert participants are mostly employed as a control variable in studies or training programmes that aim to examine the effects on novice performance. However, when experts are being trained, it is necessary to carefully develop the intervention's experimental control. For instance, Sullman et al. (2015) conducted a study with a group of professional bus drivers. The drivers were divided into two groups: one group received a training intervention relating to eco-driving, while the other group received a control intervention training unrelated to eco-driving. This rendered the assumption (1) made above superfluous by design. In this study, the salient dependent variable investigated was fuel economy, a metric that was extracted from the simulator and later measured in real life as an on-the-job efficiency. Another example of transferable metrics is found in Pagnussat et al. (2020) where forestry harvesting was trained over a duration of 40 h. The study collected data on time, tree falling direction, and cutting height during early performance testing in order to forecast performance at the conclusion of the training. Here the second assumption is met, and the third is feasible to investigate subsequently. In contrast, some relations between outcome competency and the observed metrics of training must be taken at face value, such as the study of Bratko et al. (2020) training Ukrainian border guards towards a NATO-doctrine force-on-force action.

In competency-based education, the outcome competency should be defined and standardised before training is organised (Frank et al., 2010). Following the purpose of Fig. 1, the outcome competency could be defined through more than isolated KPIs with single metrics. The closest approximation to this in this group of assessments is the Sullman et al.

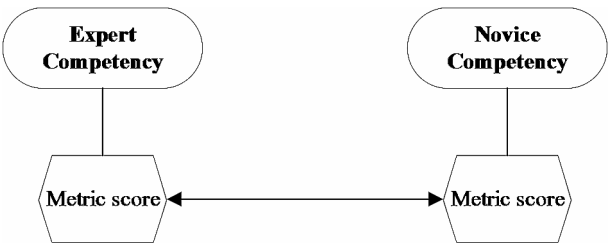

Fig. 2 Between-group comparison of raw observable metrics

(2015) study, which assessed both eco-driving knowledge and eco-driving performance but did not connect these two or other factors to describe eco-driving competency. In the next section, we discuss assessments that incorporate deeper consideration of the input data used for assessment.

4.2.2 Performance Indicator Assessments

The quality actions in instrumentation found in this group will be discussed in the following order: (1) aggregated indicators from metrics, (2) global rating scales, (3) repeated measures, and group comparison analyses. Fourteen papers in the review sample included this type of assessment, as indicated in Table 7. In contrast to the studies discussed in the previous section, the distinguishing feature of this category is the utilisation of assessment tools that are organised based on *PI* or *KPI* levels. All the studies examined skill-based outcomes, with eleven studies specifically addressing the training of medical professionals in different physician specialisations. The simulator equipment utilised consisted of 3D desktop virtual reality (VR) systems with a sample size of 10, along with two full mission concepts and two sensorized designs. Six studies utilised a between-group design, namely Bruncкhorst et al. (2015), Chang et al. (2016), Chowriappa et al. (2013), Konge et al. (2013), Poursartip et al. (2018), and Rosenthal et al. (2015). Five studies employed a mixed design namely Bajka et al. (2010), Boyle et al. (2011), Bube et al. (2019), Madsen et al. (2014), and Scavone et al. (2006). Three applied within-group design, namely de Winter et al. (2009), Duarte et al. (2013), and Liu et al. (2020).

Aggregated indicators. Modern technology frequently allows the automated generation and collection of data. In computer-based simulation, there are several chances to use this data as measurements. These measures can serve as the basis for constructing unbiased indicators for the assessment instruments. Conversely, simulators designed for specialised medical training sometimes include pre-programmed training scenarios with predetermined grading systems. To avoid this, Bube et al. (2019), Konge et al. (2013) and Madsen et al. (2014) collected all available metrics in their novice/expert between-group studies and identified which metrics discriminated between the expertise levels with statistical significance. Then the retaining metrics were aggregated into rescaled variables, which were used as simulator PI scores. No definitive mathematical formulation for the aggregation of data was identified. However, it was noted that the range of the aggregated result might be taken into account, as the input may consist of raw data with varying units and ranges. In contrast, Rosenthal et al. (2015) contend that only a single discriminant measure should be employed to express its PI, as depicted in Fig. 3. Turning to within-group design, the study of de Winter et al. (2009) gives an example of the aggregation of simulator metrics from a large dataset. Data related to how quickly the 2578 trainees executed tasks in driver training contained 209 different driving tasks and thirty-three different training blocks. Variables with a standard deviation less than 1 and redundant variables with intercorrelations above 0.8 were excluded, resulting in 457 variables. Next, a 2578×457 matrix of raw data was z-transformed, and the Pearson correlations of the matrix were submitted to an exploratory factor analysis, detecting one factor explaining 9% of the variance. This factor was identified as a latent pattern and supported by the scree plot and an eigenvalue of 150. From the z-transformed data of the identified factor, an operational score was aggregated using the Bartlett method.

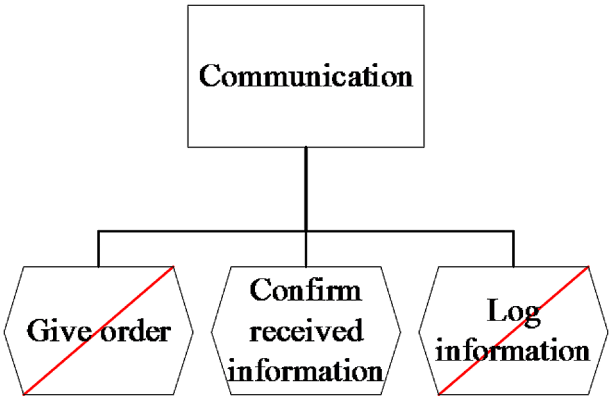

Fig. 3 Multiple metrics with discriminatory ability of expertise is either aggregated or individually transformed into a unitary variable representing the performance indicator

Global rating scales (GRS) can be viewed as a KPI-level instrument in the Fig. 1 structure and are used as assessment and feedback instruments in several of the studies. These instruments are either derived from known methods for assessing performance through observation, or they are designed specifically for simulator scenarios that utilise automated simulator data. Bajka et al. (2010) argue that their multi-metric scoring system should hold the core three features: (1) Relevance refers to the use of outcome-specific measurements that are clearly connected to the procedure or task. (2) Balance refers to ensuring that the metrics used to calculate the score do not contradict each other. For example, shorter time usage should not result in a high score if other indicators are insufficient. (3) Simple, in that the form should be straightforward, allowing for quick feedback based on the scoring. The comments should be concise, helpful, and provide assistance. Their solution to design this instrument was to engage two experts who performed a hierarchical task decomposition of the training scenario. Then fifteen observable metrics corresponding to the task decomposition was identified and grouped into 4 PIs. The experts then subjectively weighted the metrics' importance and scaled their score contribution. A similar approach was reported in Chowriappa et al. (2013) where they used expert consensus by the Delphi method for their task decomposition and to "*define, weight, integrate and configure*" the metrics into a hierarchical scoring system. This resulted in ten metrics over 5 PIs. To operationalize the instrument with a scale system, they tested it on an expert group to form a baseline with their mean metric scores. Each PI was then standardized with the expert means at 3.5 on a 5-point scale, with 2 and 4 at +/- 1 standard deviation of the expert group performance. The global score was then given by the sum of each PI multiplied with their assigned weight fraction. Similarly, to combine different PIs into a GRS Rosenthal et al. (2015) also propose standardization. The proposal suggests modifying the PIs using expert performance data. The means of the PIs would be standardised to a value of one hundred, with a standard deviation of twenty. Additionally, the PI scores would be multiplied by a weight fraction in order to get a global sum score. This quantifies the level of a novice's performance compared to that of a typical professional on a scale of 100 points with no upper limit. Adoption of existing assessor observation instruments was reported in Brunckhorst et al. (2015) with the use of the objective structured assessment of technical skills (OSATS), the non-technical skills for surgeons rating system (NOTSS), and the rigid ureteroscopy evaluation score (RUES). In contrast, Poursartip et al. (2018) explore a different approach to weighting metric data into a combined score. With simulator data from four levels of expertise, they hypothesise

constants to be multiplied with the metrics accumulating to the sum score. Approximations to these constants were found with optimalization methods in MATLAB using the generic algorithm function and the fmincon function.

Analyses of repeated measures and group comparisons. Utilising repeated measures in a training setting is a suitable approach for evaluating the impact of learning and improvements made. The study of Duarte et al. (2013) is an example of within-group repeated measures without a control group. The researchers utilised predetermined parameters from the simulator to evaluate the performance of novices over ten consecutive laparoscopic training sessions. Given that the number of repetitions was enough for the sample to reach their maximum performance, a Freidman's ANOVA demonstrated the presence of a learning curve. Subsequently, Dunnett's approach was employed to pairwise compare each session with the tenth in order to identify the point at which the learning curve plateaued, which was determined to occur between the fourth and fifth sessions. Lastly, non-linear regression was used to find that 4.26 repetitions were needed to achieve their target of an 80% successful performance. Learning curves were also investigated in Madsen et al. (2014) where between-group comparison was used to define the level of successful novice performance In brief, their five-step approach to assessing performance and learning curves comprises (1) identification of scenarios, (2) testing the validity of simulator scenarios and metrics, (3) assessing the reliability of simulator scenarios and metrics, (4) setting the performance standard, and (5) exploring learning curves. During step 2, both novices and experts made two attempts at the scenario, and the metrics were evaluated using a dichotomous assessment. Metrics with a pass rate of less than 50% among experts were eliminated, and only the remaining metrics with discriminatory ability were included. In step 3, the metrics were evaluated to determine the test-retest reliability between the two pilot attempts. The upper performance level was determined in step 4 based on the median of the expert group, whereas the lower pass-fail level was determined using the contrasting groups technique. The contrasting groups method plots the performance distribution of the two groups in order to identify the point at which their graphs intersect. During step 5, novices underwent repetitive training of the scenario in order to generate learning curves for the validated scoring system. Interestingly, novices required three iterations to achieve a pass/fail rate of 63% and four iterations to reach the median level of expertise, after which the progress levelled off. The antecedent study of Konge et al. (2013) reports a similar approach; however, the reliability of the scoring system and number of iterated scenarios were explored with generalizability theory using the GENOVA software. First, a 1-facet balanced G-study was run to estimate variance components. Then, the components were used in a D-study to find a reliability and generalizability coefficient of 0.67 for the aggregated score system. To reach a reliability level above 0.8 novices needed to perform four repetitions of the training procedure.

4.2.3 Composite Methods

These examples of assessment instrumentation differ from the previous two sections as they create the path from collected metrics to outcome competency by combining more complex methods. This occurred with four studies of the review sample, as seen in Table 9. Ernstsen and Nazir (2020) investigated an approach to assessment with full mission simulators that allows to integrate different KPIs in a Bayesian network of multiple PI levels and metrics.

This approach used an analytic hierarchical process to weight PIs, creating a network that can incorporate subjective and objective input from dichotomous or scaled observational metrics. In a training transfer study, Rauter et al. (2013) used a full-mission rowing simulator and a sensorized rowing boat to compare biomechanical performance measures and subjective trainer assessments in a training programme. Iterations of rowing movements produced data throughout the training exercises, and the coefficient of variance in these measures was summed and used to indicate consistency in performance. Rhienmora et al. (2011) investigated data from a 3D desktop VR dental surgery simulator, where the metrics were z-scaled and the steps of the procedure were used as chains in a Hidden Markov Model (HMM). Based on applied force, instrument position, and orientation, the HMM algorithm was able to distinguish between novices and experts while providing feedback. Mackel et al. (2007) also used HMM to assess novices and experts in pelvic examination on a sensorized part task mannequin. Based on transitions between examination steps, the Markov model was able to identify group affiliation with 92.7% accuracy.

4.3 Discussion Summary

As demonstrated above, there are a variety of quality procedures applied for the development of assessment instruments and for the validity of evidence of computer-based simulation training and assessment. Only a few studies focused on the entire process of developing their assessment instruments. However, all the included studies contribute insight into what metrics could be used for measurement and how these could be used for assessment in the field, condition, scenario, and training outcome of the respective study. In comparison to the above levels of methodologies, the assessments made by *observable metric comparisons* could, by extension, be integrated into the approaches found in *the performance indicator assessments* category. As such, directly comparing raw metrics between groups, aggregating multiple metrics into performance indicators, and transforming multiple performance indicators into global rating scales are distinct levels of a similar approach. In turn, these approaches might be helpful for future studies that follow the approaches of the four studies in the *composite methods* category. However, these differ from the rest of the review pool in that they are too complex to help address our final objective; that is, they would be hard to apply to a different context without replicating the whole approach as reported. The final objective was to use the review to propose an approach to assessment instrumentation that can be applied independent of discipline, within the range of current or emerging technology, and for any focal outcome competency. This should enable an approach that could be applied even as the frames of simulator training evolves. Approximating such a generic approach is suggested with a five-step structure based on the approaches found in Konge et al. (2013), Madsen et al. (2014) and Rosenthal et al. (2015). Resulting in Table 10, these studies focused on the development of assessment instruments and were used as a baseline for our model, supplemented with the quality procedures found in the review. To guide the reader, we provide our suggested reference for each step while reminding them to contemplate the variety of quality procedures discussed in the above sections. In the context of a competency development process, the outcome competency should be priorly defined in reference to a competence requirement. Otherwise, a training needs analysis could precede step 0 in the following approach.

Table 10 5 steps to creating an assessment instrument

0	Outcome competency	Conferatur
1	**Scenario formulation and identification of metrics available.** *0.1 Hierarchical task decomposition or hierarchical task analysis to formulate procedures, checklists, or instructions of the scenario.* *0.2 Identification of automated simulator metrics, observational metrics, and self-report items.*	Chowriappa et al. (2013)
	Pilot and scoring system validation. *0.1 Pilot the scenario twice with two or more expertise groups.* *0.2 Establish content validity by including metrics with discriminatory ability and excluding metrics with high inter-item correlation.* *0.3 Decide aggregation and weight of performance indicators.* *0.4 Standardize performance indicators and create global rating scale.* *0.5 Test discriminatory ability of GRS.*	Madsen et al. (2014) de Winter et al. (2009) Chowriappa et al. (2013) Rosenthal et al. (2015)
	Reliability of scoring system. *0.1 Internal consistency of metrics.* *0.2 Test/retest reliability by correlation of metrics and GRS.*	Bube et al. (2019) Madsen et al. (2014)
4	**Performance standard.** *0.1 Expert group central tendency.* *0.2 Pass/fail level by contrasting groups method.*	Konge et al. (2013)
5	**Application and analysis of outcome competency.** *0.1 Learning curve analysis for construct validity.* *0.1 Training transfer analysis for concurrent validity.* *0.1 Regression on other dependent measure for predictive validity.* *0.2 Update scoring system index with new data.*	Duarte et al. (2013) Sullman et al. (2015) de Winter et al. (2009)

4.3.1 Limitations

Keyword search is an approach that returns a greater list of results compared to operationalizing the PISCO (Table 1) and the criteria (Table 3) into exhaustive Boolean search strings. However, with multiple databases, the individual search stings will differ in terminology and complexity, which produces unequal operationalizations. The keyword approach disseminates the search terms more equally across databases but is dependent on the keyword indexing made by the authors and the publishing journal. This indexing might be limited in number by the journal and selected by the authors to reflect the research field, purpose, and outcomes of the article rather than methodology, tools, and measurements of interest. Considering the precision of the search process (see Table 5), the observed difference in effectiveness can be attributed to the database population, author indexing, or search string. As no records from the ERIC database meet the criteria, there is perhaps a dissonance between this review's scope and the database's population.

Interrater reliability is a face validity quality for the systematic literature review protocol. To ensure this quality in the operationalization of search strings with four different databases, external consulting was utilized. The expertise of the researchers authoring this paper brings value to keeping true and transparent with the methodology of a systematic literature

review. For consistency in eligibility criteria assessment, one author performed these with support from the others.

The focus of this paper concern competency assessment with computer-based simulator training and assessment in the context of competency-based education. In effect, this delimits the scope of not considering the necessary efforts to define and establish the targeted outcome competency that the assessment instrument should intend to capture. In a full-cycle process, there would also be a need to verify that the trained competency, as acquired by training and assessed by the created instrument, varies with the real-life performance of the professional competency. Some examples of transfer measures were discussed. This constitutes the importance of the chosen analyses in step 5 of the proposed approach (Table 10) and the veracity of the "experts" applied to scale and index the assessment instrument. That is to say, the expertise levels compared derive from real-life competencies. In contrast, the novice groups of the reviewed studies primarily comprise participants whose proficiency is acquired in an educational setting. Although these two demographic groups are compared in the same context and condition, it cannot ensure the transferability of novice knowledge and skills to real-life performance without first establishing external validity for the training programme and assessment instrument. Consequently, apprenticeship is still a relevant integration in professional education.

Each procedure of our Table 10 model had several alternatives, except for step 2.3 concerning the weighting of aggregated performance indicators. In the review pool, this procedure was either decided by the author or by a consensus between a few experts. Although there are options to describe the level of such consensus, this literature review found no means for mitigating subjectivity in this operation. Another annotation is the lack of guidance on how aggregation should be mathematically formulated.

4.3.2 Future Research

As previously stated, the theoretical exercises presented in this work should be supplemented with an empirical investigation. A new study has been launched to utilise the methods of the suggested model in order to create evaluation tools that can be used across multiple simulator technologies and that can be applied in real-world situations. Collecting evidence to support the validity arguments will be essential in order to ensure the relevance and effectiveness of the instruments.

By broadening the utilisation of competency-based assessment tools across many fields and examining their efficacy across diverse domains, we can gain more profound insights. This could help us understand how competencies are developed in a single field and how they could be transferred to another efficiently. However, we are not able to close the gap on interdisciplinary and cross-domain assessment instrumentation alone. Assessment instrumentation that fits training outcomes regardless of professional discipline is an issue to be revisited every time one variable of the equation changes, whether it is evolving technologies, differing populations, or novel applications of simulators in competency development.

Additional theoretical research, grounded in empirical evidence, could assist professionals in the field of training research in discussing optimal methodologies. The range of choices for instrumentation may appear limitless when considering all the different disciplines and domains. However, in the preceding section, we identified several aspects that need additional exploration.

5 Conclusion

This systematic literature review has scoped across different fields and collected an overview of recent research with assorted approaches to assessment instrumentation. The field of medical training dominated as the major contributor to this sample (40%), although computer-based simulation was found to be applied in a broad variety of fields. We aimed to answer how competency is assessed with computer-based simulation. Objectives were to (1) provide a theoretical frame to operationalize components of assessment instruments, which was done through the two dimensions of outcome competency and component level. Then, (2) performing the systematic literature review, which extracted data through these dimensions and described the characteristics of the studies. Competency was found to be assessed with a focus on skill-based outcomes within a range of modern simulator technologies, from 2D desktop interfaces to advanced 3D and virtual reality integrations. Further, quality procedures were highlighted and then synthesised to (3) propose a generic approach to new assessment instrumentation based on the quality characteristics found. The proposed approach could be helpful to leverage the frames of current simulator technologies and their integration in training and education programmes. By allowing the integration of varied factors from measurement sources within and besides the digital environment, future applications of simulators could evolve either towards more automated assessment or instructor observation. This is worth considering when training and assessing professional competency, which might require considering multiple outcome dimensions within the spectrum of cognitive-, skill-based-, and affective outcomes. The proposed approach could help researchers and practitioners maintain the integrity of validity considerations throughout the competency-based education cycle, from the starting point of competence requirements through scenario formulation and practice through computer-based simulator training, followed by assessment of specific professional competencies. Finally, evidence collected from this process could help evaluate how the obtained competencies correspond with the requirements of established competence standards. This work can contribute to fields of research and practice that are applying modern technologies for educational assessment by challenging the conduct of assessment instrumentation with an emphasis on quality procedures. The authors welcome debate and the continuation of work along this track as the domain further matures.

Funding Open access funding provided by University Of South-Eastern Norway

References

Bajka, M., Tuchschmid, S., Fink, D., Szekely, G., & Harders, M. (2010). Establishing construct validity of a virtual-reality training simulator for hysteroscopy via a multimetric scoring system [Article]. *Surgical Endoscopy*, *24*(1), 79–88. https://doi.org/10.1007/s00464-009-0582-4.

Boyle, E., O'Keeffe, D. A., Naughton, P. A., Hill, A. D., McDonnell, C. O., & Moneley, D. (2011). The importance of expert feedback during endovascular simulator training. *Journal of Vascular Surgery*, *54*(1), 240–248e241. https://doi.org/10.1016/j.jvs.2011.01.058.

Bratko, A., Hashchuk, V., Suslov, T., Misheniuk, R., Zhuravel, V., & Havryliuk, V. (2020). Assessing the effectiveness of Tactical skills Level when using a laser tag type two-way Skirmish Simulator. *Brain-Broad Research in Artificial Intelligence and Neuroscience*, *11*(1), 189–203. https://doi.org/10.18662/brain/11.1/23.

Brunckhorst, O., Shahid, S., Aydin, A., McIlhenny, C., Khan, S., Raza, S. J., Sahai, A., Brewin, J., Bello, F., Kneebone, R., Khan, M. S., Dasgupta, P., & Ahmed, K. (2015). Simulation-based ureteroscopy skills training curriculum with integration of technical and non-technical skills: A randomised controlled trial [Article]. *Surgical Endoscopy*, *29*(9), 2728–2735. https://doi.org/10.1007/s00464-014-3996-6.

Bube, S. H., Hansen, R. B., Dahl, C., Konge, L., & Azawi, N. (2019). Development and validation of a simulator-based test in transurethral resection of bladder tumours (TURBEST). *Scand J Urol*, *53*(5), 319–324. https://doi.org/10.1080/21681805.2019.1663921.

Carraccio, C., Wolfsthal, S. D., Englander, R., Ferentz, K., & Martin, C. (2002). Shifting paradigms: From Flexner to competencies. *Academic Medicine*, *77*(5), 361–367. https://doi.org/10.1097/00001888-200205000-00003.

Chang, J., Banaszek, D. C., Gambrel, J., & Bardana, D. (2016). Global rating scales and Motion Analysis are Valid Proficiency Metrics in virtual and benchtop knee arthroscopy simulators [Article]. *Clinical Orthopaedics and Related Research*, *474*(4), 956–964. https://doi.org/10.1007/s11999-015-4510-8.

Chetlen, A. L., Mendiratta-Lala, M., Probyn, L., Auffermann, W. F., DeBenedectis, C. M., Marko, J., Pua, B. B., Sato, T. S., Little, B. P., Dell, C. M., Sarkany, D., & Gettle, L. M. (2015). Conventional Medical Education and the history of Simulation in Radiology [Review]. *Academic Radiology*, *22*(10), 1252–1267. https://doi.org/10.1016/j.acra.2015.07.003.

Chowriappa, A. J., Shi, Y., Raza, S. J., Ahmed, K., Stegemann, A., Wilding, G., Kaouk, J., Peabody, J. O., Menon, M., Hassett, J. M., Kesavadas, T., & Guru, K. A. (2013). Development and validation of a composite scoring system for robot-assisted surgical training–the robotic skills Assessment score [Article]. *Journal of Surgical Research*, *185*(2), 561–569. https://doi.org/10.1016/j.jss.2013.06.054.

Colombo, S., & Golzio, L. (2016). The Plant Simulator as viable means to prevent and manage risk through competencies management: Experiment results [Article]. *Safety Science*, *84*, 46–56. https://doi.org/10.1016/j.ssci.2015.11.021.

Cook, D. A., Zendejas, B., Hamstra, S. J., Hatala, R., & Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Health Sci Educ Theory Pract*, *19*(2), 233–250. https://doi.org/10.1007/s10459-013-9458-4.

Dawe, S. R., Pena, G. N., Windsor, J. A., Broeders, J. A., Cregan, P. C., Hewett, P. J., & Maddern, G. J. (2014). Systematic review of skills transfer after surgical simulation-based training. *British Journal of Surgery*, *101*(9), 1063–1076. https://doi.org/10.1002/bjs.9482.

de Winter, J. C., de Groot, S., Mulder, M., Wieringa, P. A., Dankelman, J., & Mulder, J. A. (2009). Relationships between driving simulator performance and driving test results [Article]. *Ergonomics*, *52*(2), 137–153. https://doi.org/10.1080/00140130802277521.

Duarte, R. J., Cury, J., Oliveira, L. C., & Srougi, M. (2013). Establishing the minimal number of virtual reality simulator training sessions necessary to develop basic laparoscopic skills competence: Evaluation of the learning curve. *International Braz J Urol : Official Journal of the Brazilian Society of Urology*, *39*(5), 712–719. https://doi.org/10.1590/S1677-5538.IBJU.2013.05.14.

Ebnali, M., Hulme, K., Ebnali-Heidari, A., & Mazloumi, A. (2019). How does training effect users' attitudes and skills needed for highly automated driving? *Transportation Research Part F: Traffic Psychology and Behaviour*, *66*, 184–195. https://doi.org/10.1016/j.trf.2019.09.001.

Ernstsen, J., & Nazir, S. (2020). Performance assessment in full-scale simulators - a case of maritime pilotage operations [Article]. *Safety Science*, *129*, 104775. https://doi.org/10.1016/j.ssci.2020.104775.

Evans, K., & Kersh, N. (2014). Training and Workplace Learning. In *The Wiley Blackwell Handbook of the Psychology of Training, Development, and Performance Improvement* (pp. 50–67). https://doi.org/10.1002/9781118736982.ch4.

Flin, R., & O'Connor, P. (2017). *Safety at the Sharp End*.

Frank, J. R., Mungroo, R., Ahmad, Y., Wang, M., De Rossi, S., & Horsley, T. (2010). Toward a definition of competency-based education in medicine: A systematic review of published definitions. *Medical Teacher*, *32*(8), 631–637. https://doi.org/10.3109/0142159X.2010.500898.

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, *26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x.

Grossman, R., Heyne, K., & Salas, E. (2014). Game- and Simulation-Based Approaches to Training. In *The Wiley Blackwell Handbook of the Psychology of Training, Development, and Performance Improvement* (pp. 205–223). https://doi.org/10.1002/9781118736982.ch12.

Hjellvik, S., & Mallam, S. (2021). *Adaptive training with cloud-based simulators in maritime education* Proceedings of the International Maritime Lecturers' Association. Seas of transition: setting a course for the future, https://commons.wmu.se/cgi/viewcontent.cgi?article=1019&context=imla2021

Hjelmervik, K., Nazir, S., & Myhrvold, A. (2018). Simulator training for maritime complex tasks: An experimental study. *WMU Journal of Maritime Affairs*, *17*(1), 17–30. https://doi.org/10.1007/s13437-017-0133-0.

Holmboe, E. S., Sherbino, J., Long, D. M., Swing, S. R., & Frank, J. R. (2010). The role of assessment in competency-based medical education. *Medical Teacher*, *32*(8), 676–682. https://doi.org/10.3109/0142159X.2010.500704.

Iqbal, M. U., & Srinivasan, R. (2018). Simulator based performance metrics to estimate reliability of control room operators [Article]. *Journal of Loss Prevention in the Process Industries*, *56*, 524–530. https://doi.org/10.1016/j.jlp.2017.10.011.

Konge, L., Annema, J., Clementsen, P., Minddal, V., Vilmann, P., & Ringsted, C. (2013). Using virtual-reality simulation to assess performance in endobronchial ultrasound [Article]. *Respiration*, *86*(1), 59–65. https://doi.org/10.1159/000350428.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, *78*(2), 311–328. https://doi.org/10.1037/0021-9010.78.2.311.

Li, G. Y., Mao, R. Z., Hildre, H. P., & Zhang, H. X. (2020). Visual attention Assessment for Expert-in-the-Loop Training in a Maritime Operation Simulator. *IEEE Transactions on Industrial Informatics*, *16*(1), 522–531. https://doi.org/10.1109/Tii.2019.2945361.

Liu, Y. S., Lan, Z. R., Cui, J., Krishnan, G., Sourina, O., Konovessis, D., Ang, H. E., & Mueller-Wittig, W. (2020). Psychophysiological evaluation of seafarers to improve training in maritime virtual simulator [Article]. *Advanced Engineering Informatics*, *44*, 101048. https://doi.org/10.1016/j.aei.2020.101048.

Loukas, C., Nikiteas, N., Kanakis, M., & Georgiou, E. (2011). Deconstructing laparoscopic competence in a virtual reality simulation environment [Article]. *Surgery*, *149*(6), 750–760. https://doi.org/10.1016/j.surg.2010.11.012.

Mackel, T. R., Rosen, J., & Pugh, C. M. (2007). Markov model assessment of subjects' clinical skill using the E-Pelvis physical simulator [Article]. *Ieee Transactions on Biomedical Engineering*, *54*(12), 2133–2141. https://doi.org/10.1109/tbme.2007.908338.

Madsen, M. E., Konge, L., Norgaard, L. N., Tabor, A., Ringsted, C., Klemmensen, A. K., Ottesen, B., & Tolsgaard, M. G. (2014). Assessment of performance measures and learning curves for use of a virtual-reality ultrasound simulator in transvaginal ultrasound examination. *Ultrasound in Obstetrics and Gynecology*, *44*(6), 693–699. https://doi.org/10.1002/uog.13400.

Manuel, M. E. (2017). Vocational and academic approaches to maritime education and training (MET): Trends, challenges and opportunities. *WMU Journal of Maritime Affairs*, *16*(3), 473–483. https://doi.org/10.1007/s13437-017-0130-3.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, *65*(9 Suppl), 63–67. https://doi.org/10.1097/00001888-199009000-00045.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*, *339*(jul21 1), b2535–b2535. https://doi.org/10.1136/bmj.b2535.

Nisizaki, C., Okazaki, T., Yoshino, R., Takaseki, R., & Murai, K. (2017). A study on evaluation method for ship maneuvering simulator training. 2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017.

O'Donnell, E., Lawless, S., Sharp, M., & Wade, V. P. (2015). A review of personalised E-Learning. *International Journal of Distance Education Technologies*, *13*(1), 22–47. https://doi.org/10.4018/ijdet.2015010102.

Ojados Gonzalez, D., Martin-Gorriz, B., Ibarra Berrocal, I., Macian Morales, A., Salcedo, A., G., & Hernandez, M., B (2017). Development and assessment of a tractor driving simulator with immersive virtual reality for training to avoid occupational hazards [Article]. *Computers and Electronics in Agriculture*, *143*, 111–118. https://doi.org/10.1016/j.compag.2017.10.008.

Okazaki, T., & Ohya, M. (2012). 14–17 Oct. 2012). A study on situation awareness of marine pilot trainees in crowded sea route. 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC).

Pagnussat, M., Hauge, T., Lopes, E. D., de Almeida, R. M. M., & Naldony, A. (2020). Bimanual Motor Skill in Recruitment of Forest Harvest Machine Operators [Article]. *Croatian Journal of Forest Engineering*, *41*(1), 25–33. https://doi.org/10.5552/crojfe.2020.623.

Passmore, J., & Velez, M. J. (2014). Training Evaluation. In K. Kraiger;, J. Passmore;, S. Malcezzi;, & N. R. d. Santos; (Eds.), *The Wiley Blackwell Handbook of the Psychology of Training, Development, and Performance Improvement* (pp. 136–153). https://doi.org/10.1002/9781118736982.ch8.

Poursartip, B., LeBel, M. E., Patel, R. V., Naish, M. D., & Trejos, A. L. (2018). Analysis of Energy-Based Metrics for Laparoscopic skills Assessment. *Ieee Transactions on Biomedical Engineering*, *65*(7), 1532–1542. https://doi.org/10.1109/TBME.2017.2706499.

Prohn, M. J., & Herbig, B. (2020). Evaluating the effects of a simulator-based training on knowledge, attitudes and driving profiles of German ambulance drivers [Article]. *Accident Analysis and Prevention*, *138*., 105466, Article 105466. https://doi.org/10.1016/j.aap.2020.105466.

Randolph, J. J. (2009). A guide to writing the dissertation literature review. *Practical Assessment, Research & Education*, *14*(13), 1–13. https://doi.org/10.7275/b0az-8t74.

Rauter, G., Sigrist, R., Koch, C., Crivelli, F., van Raai, M., Riener, R., & Wolf, P. (2013). Transfer of complex skill learning from virtual to real rowing. *Plos One*, *8*(12), e82145. https://doi.org/10.1371/journal.pone.0082145.

Rhienmora, P., Haddawy, P., Suebnukarn, S., & Dailey, M. N. (2011). Intelligent dental training simulator with objective skill assessment and feedback [Article]. *Artificial Intelligence In Medicine*, *52*(2), 115–121. https://doi.org/10.1016/j.artmed.2011.04.003.

Rosenthal, R., von Websky, M. W., Hoffmann, H., Vitz, M., Hahnloser, D., Bucher, H. C., & Schafer, J. (2015). How to report multiple outcome metrics in virtual reality simulation [Article]. *EUROPEAN SURGERY-ACTA CHIRURGICA AUSTRIACA*, *47*(4), 202–205. https://doi.org/10.1007/s10353-015-0327-7.

Ryall, T., Judd, B. K., & Gordon, C. J. (2016). Simulation-based assessments in health professional education: A systematic review. *J Multidiscip Healthc*, *9*, 69–82. https://doi.org/10.2147/JMDH.S92695.

Scavone, B. M., Sproviero, M. T., McCarthy, R. J., Wong, C. A., Sullivan, J. T., Siddall, V. J., & Wade, L. D. (2006). Development of an objective scoring system for measurement of resident performance on the human patient simulator. *Anesthesiology*, *105*(2), 260–266. https://doi.org/10.1097/00000542-200608000-00008.

Schardt, C., Adams, M. B., Owens, T., Keitz, S., & Fontelo, P. (2007). Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Med Inform Decis Mak*, *7*, 16. https://doi.org/10.1186/1472-6947-7-16.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039.

Spence, J. R., & Baratta, P. L. (2014). Performance Appraisal and Development. In *The Wiley Blackwell Handbook of the Psychology of Training, Development, and Performance Improvement* (pp. 437–461). https://doi.org/10.1002/9781118736982.ch22.

Sportillo, D., Paljic, A., & Ojeda, L. (2019). 11–14 March 2019). On-Road Evaluation of Autonomous Driving Training. 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI).

Sullman, M. J. M., Dorn, L., & Niemi, P. (2015). Eco-driving training of professional bus drivers – does it work? [Article]. *Transportation Research Part C: Emerging Technologies*, *58*(PD), 749–759. https://doi.org/10.1016/j.trc.2015.04.010.

Sutton, A., Clowes, M., Preston, L., & Booth, A. (2019). Meeting the review family: Exploring review types and associated information retrieval requirements. *Health Information & Libraries Journal*, *36*(3), 202–222. https://doi.org/10.1111/hir.12276.

Taylor, G. S., & Barnett, J. S. (2010). Training effectiveness of wearable and desktop simulator interfaces. 54th Human Factors and Ergonomics Society Annual Meeting 2010, HFES 2010, San Francisco, CA.

Taylor, G. S., & Barnett, J. S. (2013). Evaluation of wearable simulation interface for military training [Article]. *Human Factors*, *55*(3), 672–690. https://doi.org/10.1177/0018720812466892.

van Hove, P. D., Tuijthof, G. J., Verdaasdonk, E. G., Stassen, L. P., & Dankelman, J. (2010). Objective assessment of technical surgical skills. *British Journal of Surgery*, *97*(7), 972–987. https://doi.org/10.1002/bjs.7115.

Verstappen, V. J., Pikaar, E. N., & Zon, R. G. (2022). Assessing the impact of driver advisory systems on train driver workload, attention allocation and safety performance [Article]. *Applied Ergonomics*, *100*, 103645., Article 103645. https://doi.org/10.1016/j.apergo.2021.103645.

Wagner, N., Fahim, C., Dunn, K., Reid, D., & Sonnadara, R. R. (2017). Otolaryngology residency education: A scoping review on the shift towards competency-based medical education. *Clinical Otolaryngology*, *42*(3), 564–572. https://doi.org/10.1111/coa.12772.

Authors and Affiliations

Simen Hjellvik[1] · Steven Mallam[1] · Marte Fannelø Giskeødegård[2] · Salman Nazir[1]

✉ Simen Hjellvik
Simen.Hjellvik@usn.no

[1] Department of Maritime Operations, University of South-Eastern Norway, Campus Vestfold, Raveien 215, Borre 3184, Norway

[2] Department of Ocean Operations and Civil Engineering, NTNU, Larsgårdsvegen 2, postboks 1517, Ålesund 6025, Norway