University of South-Eastern Norway
Faculty of Technology, Natural Sciences and Maritime Studies

Doctoral dissertation no. 64
2020

Jørgen Ernstsen

Reducing the subjective impact in maritime simulator assessment

A performance assessment tool for maritime pilotage operations

University of South-Eastern Norway
NTNU Norwegian University of Science and Technology
UiT Norges arktiske universitet
Western Norway University of Applied Sciences

Jørgen Ernstsen

Reducing the subjective impact in maritime simulator assessment

A performance assessment tool for maritime pilotage operations

A PhD dissertation in
Nautical operations

University of South-Eastern Norway
NTNU Norwegian University of Science and Technology
UiT Norges arktiske universitet
Western Norway University of Applied Sciences

University of South-Eastern Norway
Faculty of Technology, Natural Sciences and Maritime Studies
Department of Maritime Operations
Horten, 2020

Norwegian University of Science and Technology
Faculty of Engineering
Department of Ocean Operations and Civil Engineering

UiT The Arctic University of Norway
Faculty of Science and Technology
Department of Technology and Safety

Western Norway University of Applied Sciences
Faculty of Business Administration
and Social Sciences
Department of Maritime Studies

Doctoral dissertations at the University of South-Eastern Norway no. 64
ISSN: 2535-5244 (print)
ISSN: 2535-5252 (online)

ISBN: 978-82-7860-423-6 (print)
ISBN: 82-7860-425-8 (online)

Print: University of South-Eastern Norway
Cover Photo: Jørgen Ernstsen

Dedication

For my moms and dads,

thank you for inspiring me to live and grow.

Acknowledgements

This adventure has been a privilege. It has enabled me to grow and develop in ways that I could not foresee before embarking on to the scientific vessel. It has brought me into contact with remarkable people that have supported me all the way in my journey, for which I will be forever grateful.

Firstly, I want to send my appreciations to my supervisors, Dr. Salman Nazir, Dr. Kjell-Ivar Øvergård and Dr. Bjarte Knappen Røed. Thank you for your invaluable guidance throughout this process. I have learned a tremendous amount by working with you and I will always remember the opportunities that you have provided for me.

To my friends and colleagues at the Department of Maritime Operations and in the Training and Assessment Research Group (TARG). Thank you for making it a pleasure coming into work every day. You have truly enriched my experience. Steven, Amit, Sathiya, Tae, Laura, Anne, Atle, Tine, Øyvind, Lars Christian, Monica, Thærie, Morten, Per, Per-Eirik, Thomas, Paul Nikolai, Haakon, Petter, Dmitrii, Mariia, Karina, Halvor, Mehdi, Veronica, Anne Kari, Monica, as well as all the research assistants who have been working with TARG through the years. Thank you.

I would also like to extend my formal appreciation to external individuals who have made it possible to carry out my research. Thank you to the Norwegian Coastal Administration for providing me with pilots and giving me pilotage experiences. Thank you to all the captains who spent their “free time” on helping me with research. Thank you to Chalmers University for letting me run experiments at their location. And thank you to Prof. Kluge and her research group at the Ruhr-University Bochum for their collaboration, and Prof. Veitch and Dr. Musharraf at the Memorial University of Newfoundland for their instrumental methodological support.

A special thank you to my colleagues Morten and Haakon who have given me exceptional assistance in using the full-scale simulators and in running the final experiments. I could not have done it without you. Also, Steven, I am so grateful for your time reviewing the thesis and for your general support throughout the project.

To the library at the University of South-Eastern Norway. Thank you for having the patience to help me with all my enquiries. You have been an incredible resource for my journey.

To my family and friends. Your encouragement has been more important than you perhaps know. All your words have profoundly helped on carrying on this journey. I am humble for having you all in my life.

To my life partner and best friend, Elise. I am blessed and eternally grateful for all the backing and support that you have given me throughout these years. Thank you for everything.

Abstract

Maritime navigation involves the process of monitoring and controlling the movement of a ship from one position to another. When the ship is approaching or leaving a port, a local navigational expert, the pilot, is often provided for assisting the bridge team to safely and efficiently navigate the littoral waters. However, statistics concerning maritime accidents continues to associate human errors with accidents, which advocates a scrutiny of maritime education and training (MET) for pilotage operations.

The pilotage operation is a unique phase of the voyage in which the bridge team has to rely on an external expert with whom they have no prior working experience with. Also, the pilot-bridge team often has to perform immediately, and the consequences of not performing can be calamitous. These aspects suggest a need for dedicated research on how to ensure that bridge teams have the necessary competencies for carrying out safe and efficient pilotage operations. Training and assessment are quintessential instruments for this mission.

MET facilities increasingly rely on full-scale simulators for their benefits in training for complex operations. Meanwhile, the assessment of training performance is often based on subjective criteria, which has implications on the reliability and validity of the assessment. Consequently, lacking a reliable and valid assessment method could have repercussions on the educational quality, as well as subsequent developments to the MET facilities' training and education programs.

The aim of this research is to understand and to advance performance assessment related to the use of full-scale maritime simulators in MET, with a main objective to reduce the subjective impact that can be present in the performance assessment of pilotage operations.

This thesis presents a computer-assisted assessment tool based on a structural probabilistic network, in which the assessment criteria are weighted by using an analytical hierarchical process. This tool is designed for flexibility, so it can easily be employed across a multitude of pilotage scenarios, and to be less subjective so that it can present more reliable and valid performance information to its stakeholders.

The results show that the presented assessment tool can be used for higher reliability in the assessment of technical performance, but that more research is required to assess teamwork performance reliably. The tool’s content validity is considered adequate, and that studies over time are required to assess its criterion validity effectively. Regardless, the tool is deemed opportune for generating precise and accurate assessment of training performance and could serve as a steppingstone to objective assessment.

Accurate and precise assessment of training performance are imperative for stakeholders that make executive decisions concerning the development of training programs, competency mappings of the workforce, as well as for the trainee to know his or her strengths and weaknesses in the operation. Providing executive stakeholders with information for making decisions that are based on objective performance assessment data could serve as a piece of the puzzle in the mission to reduce human errors in maritime shipping.

Keywords: Maritime; Performance Assessment; Training; Pilotage Operations; Human Factors; Full-scale Simulators; Maritime Education and Training

List of appended articles

Article 1:

Ernstsen, J., & Nazir, S. (2018). Consistency in the development of performance assessment methods in the maritime domain. *WMU Journal of Maritime Affairs*, *17*(1), 71-90. doi: 10.1007/s13437-018-0136-5

Article 2:

Ernstsen, J., Musharraf, M., Mallam, S. C, Nazir, S., & Veitch, B. (2018). Bayesian Network for Assessing Performance in Complex Navigation-A Conceptual Model. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 62*(1), 1751-1755. Sage CA: Los Angeles, CA: SAGE Publications. doi: 10.1177/1541931218621396

Article 3:

Ernstsen, J., & Nazir, S. (2019). Exploring teamwork in maritime pilotage operations. *Ergonomics (in second review)*

Article 4:

Ernstsen, J., & Nazir, S. (2019). Performance assessment in full-scale simulators – a case of maritime pilotage. *Safety Science (in review)*

List of relevant publications

Ernstsen, J., Nazir, S., Røed, B. K., & Manca, D. (2016). Systemising performance indicators in the assessment of complex sociotechnical systems. *Chemical Engineering Transactions, 53*, 187-192. doi: 10.3303/CET1653032

Ernstsen, J., Nazir, S., & Røed, B. K. (2017). Human reliability analysis of a pilotage operation. In *Safety of Sea Transportation: Proceedings of the 12th International Conference on Marine Navigation and Safety of Sea Transportation* (TransNav 2017), June 21-23, 2017, Gdynia, Poland (p. 295-300). CRC Press

Ernstsen, J., Musharraf, M., & Nazir, S. (2018). Bayesian Model of Operator Challenges in Maritime Pilotage. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 62*(1), 1813-1817. Sage CA: Los Angeles. doi: 10.1177/1541931218621411

Ernstsen, J., & Nazir, S. (2018). Human error in pilotage operations. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, *12*(1), 49-56. doi: 10.12716/1001.12.01.05

Mallam, S. C., Nazir, S., Renganayagalu, S. K., **Ernstsen, J.,** Veie, S., & Edwinson, A. E. (2018). Design of Experiment Comparing Users of Virtual Reality Head-Mounted Displays and Desktop Computers. *In Congress of the International Ergonomics Association, 822*, 240-249. Springer, Cham. doi: 10.1007/978-3-319-96077-7_25

Sharma, A., Nazir, S., & **Ernstsen, J.** (2019). Situation awareness information requirements for maritime navigation: A goal directed task analysis. *Safety Science, 120*, 745-752. doi: 10.1016/j.ssci.2019.08.016

Ernstsen, J., Mallam, S. C., & Nazir, S. (to be published). Incidental Memory Recall in Virtual Reality: An Empirical Investigation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*

Mallam, S. C., **Ernstsen, J.,** & Nazir, S. (to be published). Safety in Shipping: Investigating Safety Climate in Norwegian Maritime Workers. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

Renganayagalu, S. K., Mallam, S. C., Nazir, S., **Ernstsen, J.,** Haavardtun, P. (to be published). Impact of simulation fidelity on student self-efficacy and perceived skill development in engine room simulators. In *Safety of Sea Transportation: Proceedings of the 12th International Conference on Marine Navigation and Safety of Sea Transportation*

Summary of appended articles

Article 1

Citation: Ernstsen, J., & Nazir, S. (2018). Consistency in the development of performance assessment methods in the maritime domain. *WMU Journal of Maritime Affairs*, *17*(1), 71-90. doi: 10.1007/s13437-018-0136-5

Short summary: This article examined the consistent use of assessment throughout four major segments in the maritime industry. It reports a systematic review of literature for their methods of developing the assessment methods. The review identified whether the published research developed or based their assessment framework on robust measures.

Contribution to thesis: The systematic literature review was necessary to gain an understanding of the assessment needs in the maritime industry. This knowledge and information played a central role in the development of the assessment concept (stage 2).

Article 2

Citation: Ernstsen, J., Musharraf, M., Mallam, S. C, Nazir, S., & Veitch, B. (2018). Bayesian Network for Assessing Performance in Complex Navigation- A Conceptual Model. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 62*(1), 1751-1755. Sage CA: Los Angeles, CA: SAGE Publications. doi: 10.1177/1541931218621396

Short summary: This article presents a conceptual framework for assessing performance in complex navigation. The concept is based on the use of a Bayesian network for structuring the assessment of such operations. The article also provides necessary demarcations of the maritime industry and elements of its complexity.

Contribution to thesis

The article reports the work conducted in Stage 2 of the doctoral project which, aimed to develop a conceptual framework for performance assessment. In addition to the conceptual assessment framework, the work on demarcating the complexity of maritime navigation was integral to specifying the scope of the final assessment tool.

Article 3

Citation:

Ernstsen, J., & Nazir, S. (2019). Exploring teamwork in maritime pilotage operations. *Ergonomics (in second review)*

Short summary:

This article examines the relevance of four core teamwork factors for the case of maritime pilotage operations. Comprehensive interviews with pilots and captains were carried out, and a content analysis was conducted to explore how the teamwork factors could be applied to pilotage operations.

Contribution to thesis:

The work reported in this article was a necessary addition to the development of the assessment tool. As discussed in later chapters of this thesis, the technical requirements of pilotage operations are well-covered from both a research and practical perspective. Investigating how the core teamwork factors apply to the case of maritime pilotage operations were important to develop the teamwork dimension of the assessment tool.

Article 4

Citation: Ernstsen, J., & Nazir, S. (2019). Performance assessment in full-scale simulators – a case of maritime pilotage. *Safety Science (in review)*

Short summary: In this article, the assessment tool's reliability and validity were examined using a maritime simulator experiment. A pre-recorded pilotage scenario was developed using actors in a full-scale simulator. The scenario was then assessed by raters, in which half were assessing the scenario using conventional assessment methods, whereas the other half was using the assessment tool.

Contribution to thesis: Reliability and validity are integral to any measurement. Investigating these aspects for the assessment tool was necessary to know how accurate and consistent it measures what is intended. Finally, as the purpose of the doctoral research was to reduce the subjective impact in the performance assessment of pilotage operations in full-scale simulators, a comprehensive study on the assessment tool's reliability and validity was necessary.

List of tables

List of figures

Table of contents

Papers omitted from online publication due to publisher's restrictions

1 Introduction

The thesis' domain of interest, research aim and objectives, and research epistemology are presented in this chapter.

1.1 Maritime shipping

The maritime shipping domain is massive and increasingly complex: It is responsible for transporting ninety per cent of the global cargo, making it the most extensive means of transport. It has had this entitlement for the past 5000 years (Stopford, 2009) and has been ever-expanding since then. Maritime shipping in this context concerns the waterborne transport of passengers and cargo via navigable waterways. It has had profound implications on the world and society as we know it (Paine, 2014), and is contemporary one of the most global and economically important industries in the world (Progoulaki & Roe, 2011).

The biggest boom in maritime shipping has been witnessed in the past 70 years in parallel with economic growth, industrialisation and containerisation. The world fleet has doubled in the past decade to roughly 1.7 billion dead-weight tonnage, 85,000 ships, and around 1.5 million seafarers (BIMCO, 2015; Lane, Obando-Rojas, & Wu, 2002). This fleet includes oil tankers, bulk carriers, general cargo ships, container ships, gas carriers, chemical tankers, offshore supply ships, ferries; but excluding fishing ships, barges, military ships, yachts, and offshore rigs- and platforms. This development has undoubtedly increased the complexities and competitiveness of the domain (Bhattacharya, 2015; UNCTAD, 2018).

Furthermore, this development must be coupled with safety frameworks to ensure that the hazardous nature of shipping is not provoked, as discussed in Hetherington, Flin and Mearns (2006) and in Håvold (2005). Extensive research on the topic is continuously carried out, in which the safety and efficiency of maritime shipping are researched to identify, investigate, and innovate solutions to all the challenges that arise with its development. This perpetual evolution is critical for any competitive high-risk industries, and continuous efforts are necessary to keep assuring that maritime shipping maintains a safe and environmentally friendly way for long-distance transportation of goods and passengers (Ashmawy, 2012).

Nonetheless, accidents and incidents still occur, in which the human element issues related to the accidents have been assigned high priority by the International Maritime Organization (IMO) in the *Standard of Training, Certification and Watchkeeping for Seafarers* (STCW) Manila amendment (2010) and IMO Resolution A.947(23) (IMO, 2011). Originally, however, IMO's vision to improve safety involved the technical aspects of shipping solely; the STCW was not introduced until the 1970s (Wilcox, 2000). One reason for the STCW's introduction and development was that human errors was and is repeatedly found as the main contributors of maritime accidents (Hetherington et al., 2006; Rumawas, 2016; Wilcox, 2000). Well-known examples are the SS Torrey Canyon oil spill in 1967 and the recent Helge Ingstad accident in 2018 (Accident Investigation Board, 2018; Liberia, 1967). These examples along global maritime accident statistics underpin the significance of understanding human performance and errors in the prevention and mitigation of such accidents. Moreover, contemporary studies that have examined antecedents and potential remedies related to the high number of these accidents found that, in addition to installing proper safety management systems, maritime education and training have also a tremendous impact on maritime safety (Ashmawy, 2009; Vederhus, Ødegård, Nistad, & Håvold, 2018).

1.1.1 Using full-scale simulators in maritime education and training

Contemporary full-scale simulator training is an important element in maritime education and training, in addition to other physical resources such as classrooms, audio and visual aids, libraries, but also swimming pools, lifeboats- and fire drill equipment (Sampson, 2004). Full-scale simulators, however, enable seafarers to learn and rehearse maritime operations and procedures in safe and (relatively) cheap environments compared to real operational training. The simulators are used across many industries and throughout the maritime industry – from navigational operations to anchor handling. At the same time, research that investigates how the simulators should be used point repeatedly to a need for dedicated attention to assessment in order to ensure that simulator training complies with the STCW requirements for training and assessment (Kobayashi, 2005).

Gekara, Bloor and Sampson (2011) emphasised that the lack of structure in simulator assessment can actually be detrimental to maritime safety. One issue was conveyed by Emad and Roth (2008) and Ramsden (1997) that the assessment framework itself impacts the perception and development of learning objectives for both the students and the course administrative. One consequence, they discussed, is that the perception of the training objective orients towards passing competency tests instead of truly learning the necessary skills and knowledge. Furthermore, Sellberg (2017) reported a systematic literature review on training and assessment in simulators for bridge operations. From the analysis of all the publications between the years 2000 and 2016, the research unveiled that more research on the use of maritime simulators is needed and that we have "more questions than answers". Then, a subsequent literature review identified a need to pay attention to the development of assessment methods within the industry (Ernstsen & Nazir, 2018a). In light of this, a focus on the assessment methods for maritime bridge simulators is appropriate. Two considerations related to the quality of such methods are reliability and validity (Kimberlin & Winterstein, 2008).

1.1.2 Reliability and validity in simulator performance assessment

The trainees' performance in simulators that assessors observe consists of partly true measurement and partly measurement error. Reliability is the measurement consistency of an individual's performance on a test (Safrit & Wood, 1989). A measure will always consist of some measurement error in practice; thus, reliability can be considered as the amount of error that is deemed acceptable for an effective and practical use of the instrument (Atkinson & Nevill, 1998). Furthermore, many terms are interchangeably used to refer to reliability: e.g., repeatability, reproducibility, consistency, agreement, concordance and stability (or the antonym 'variability'). Reliability is investigated from different perspectives, e.g., between items in a questionnaire (interitem reliability), between raters (agreement), and reliability of a measurement tool (variability). A reliable measurement is important and a pre-condition for validity (Kimberlin & Winterstein, 2008).

Validity of a measurement is, in large part, about reducing the measurement error. It refers to a measurement tool's ability to achieve its intended outcomes (Atkinson & Nevill, 1998). This is of paramount importance in terms of enabling the instrument's stakeholders to trust its results (Cook & Hatala, 2016). There are many frameworks and taxonomies for validity; however, the three classic types of validity are content validity, construct validity and criterion validity (Messick, 1987). Content validity addresses how well the constructs are operationalised and must be subjectively determined by experts in the field. An important element to this, then, is being transparent on what constitutes the tool. Construct validity means to assess if the scores vary as expected and can explain variation in the construct that was operationalised. This type of validity is a judgement based on the accumulation of evidence from different studies, in which content- and criterion-related validity evidence also contributes. Lastly, criterion validity refers to the measurement's correlation with other measures of the construct, i.e., whether the test results correlate with other established tests and 'true' measures.

In summary, measurement tools that are valid and reliable is a crucial component of performance assessment in simulators as these tools provide feedback to the trainee and student, as well as to the instructor concerning the effectiveness of the training programme.

1.2 The research aim and objective

The aim of this research is to understand and advance performance assessment of navigation students and trainees in full-scale maritime simulators. This research focuses on studying navigation in pilotage operations with the purpose of developing an assessment tool that improves the reliability and validity of the performance assessment of these operations when carried out in full-scale simulators.

The main objective of this research is to *reduce the subjective impact in the performance assessment of pilotage operations*, particularly in connection to training and education in full-scale simulators, as well as reliability and validity considerations of such assessment. This objective is addressed by the following research questions:

1. What is the consistency of how performance assessment methods are developed in the maritime domain?
2. How can a performance assessment framework for navigational operations be conceptualised?
3. What are the assessment requirements for the assessment tool?
4. What is the reliability and validity of the assessment tool?

1.3 What kind of knowledge does the methodology aim to produce?

1.3.1 Epistemological considerations in scientific knowledge production

There are several perspectives on what constitutes knowledge. The perspectives have been defined and placed on a continuum from radical relativist to naïve realist (Madill, Jordan, & Shirley, 2000). The perspectives have implications for the methodological design in the research. The relativist subscribes to a view on knowledge that there is no such thing as 'pure experience', whereas the realist entails the belief that the data ought to provide information about the world and how things really are. The relativist is interested to explore cultural and discursive resources in different contexts, and suitable methods enables the researcher to unpack such resources, where data collections techniques are sensitive to contradictions, variations and tensions among data sources. In contrast, the realist aims to collect unbiased data, where the data is conveyed free and open to provide and facilitate true and undistorted representations about the world.

Ultimately, attaching labels, definitions and nomenclature is problematic; e.g., what does it mean for something to be real? Researchers should aim to not be too categorical, but rather clearly express the type of knowledge that is aimed to produce irrespective of the subscribed epistemological positions and labels (Willig, 2013). Johnson and

Onwuegbuzie (2004) advocates that research approaches should be combined in ways to best answer the research questions; which in essence rejects dogmatism by combining quantitative and qualitative research techniques, methods, approaches, concepts and language to best answer the research questions. However, Johnson and Onwuegbuzie (2004) acknowledges that the mixed-method approach to research can have the tendency to give more attention to, and perhaps be more appropriate for, applied research.

1.3.2 Epistemological considerations in the current research

The research that is reported in this thesis is applied, meaning that the purpose is to solve specific practical problems, in contrast to basic research that aims to expand the existing base of scientific knowledge. There are certain challenges with applied research that the doctoral research wants to address. Applied research, especially on highly specialised domains, often have predicaments with the population size and available sample. The basic researcher, simply speaking, has the luxury of designing and optimising experiments for answering the research question, whereas the applied research must also consider, balance and combine compatible methods in order to collect enough data to answer the research questions. One example is the use of purposive random sampling, as suggested in Singh and Masuku (2014), in the research reported in Article 4, in which random sampling is considered to provide better internal validity to the experiment design.

The challenge with sample size is aggravated by the unit-of-analysis dilemma and consequently makes team research difficult (McIntyre & Salas, 1995). Team research are aggregations of individuals and therefore the unit-of-analysis is the team, not the individuals, which makes sampling more difficult concerning the need to satisfy the statistical procedures that are used to arrive at reasonable conclusions.

The difficulties of specialised applied research and unit-of-analysis dilemma hold true for the current doctoral research. The limitations that are here addressed, thus, have implications for the methodological framework in the current thesis. Compromises and pragmatic decisions were necessary to arrive at reasonable inferences and knowledge concerning the research question. For instance, the examination of construct validity would be better served by conducting a comprehensive factor analysis, but the amount of data required rendered this method unreasonable; thus, it was necessary to instead employ mixed methods for collecting data that were practically obtainable, as suggested in Johnson and Onwuegbuzie (2004).

1.4 Thesis structure

This doctoral thesis has seven chapters in addition to the four appended articles. Chapter 1 provides an introduction to the maritime domain in the context of maritime education and training, as well as the research scope and objective, philosophical clarifications, summary of publications and the structure of the thesis. Chapter 2 introduces the reader to the background and context of the research, which consists of an introduction to maritime navigation and pilotage operations, but also a delimitation of the unit of assessment. Such as the specific details of the bridge team that is to be assessed in this current research. In chapter 3, the theoretical framework is presented. Research on training and assessment is presented, as well as the relevant theory on teamwork. Chapters 2 and 3 aim to give a foundation for the last part of the thesis that focuses on the specifics of the research reported in the four articles.

Chapter 4 presents a methodological overview and framework of the doctoral thesis and the respective stages (section 4.1) and the methodological tools that are used across the research stages and in the respective articles (section 4.2 and 4.3). Chapter 5 presents the results from the four articles, which are further discussed in chapter 6. This sixth chapter also provides reflections on the limitations surrounding this doctoral work and ends with a discussion on future research and stakeholder recommendation. The research is concluded in chapter 7 by summarising and highlighting the main contributions of this research project. Finally, the full list of references for the dissertation is presented. The references that are used in each of the appended articles are provided at the end of each article.

2 Background and context

2.1 Maritime navigation

Maritime navigation involves the process of monitoring and controlling the movement of a watercraft from one place to another. The ship's massive inertia prescribes that all navigational manoeuvres must be well-planned. Adding to this, the difficulties of inferring the intentions of surrounding ships, adjusting to unpredictable hydrodynamic patterns and time schedules contribute to making navigation a stressful and challenging task (Lee & Sanquist, 2000). Technological innovations in recent decades have eliminated some of the time-intensive, repetitive and error-prone tasks of navigation, but has also introduced a complex layer of mental scaling and transformation exercises that often manifest in stressful situations while navigating (Sharma, Nazir, & Ernstsen, 2019; Woods, Johannesen, Cook, & Sarter, 1994).

A multitude of different functions is necessary to safely and efficiently navigate a ship from port to port. They are ranging from logistics, ship construction, classification and quality inspections, training and assessment, to the front-end of engine - and bridge operations. The bridge of modern ships also serves as a common platform for alarming- and control station for the onboard machinery. The equipment and instruments onboard must be pre-approved and complying with various (IMO) standards before being installed on the bridge. An important consideration, as an example, is to ensure that electrical and electronic equipment do not interfere with electromagnetic navigation equipment during the voyage.

A ship's voyage can consist of different phases: open-ocean, coastal cruising (i.e., within 50 nautical miles of the coast), inland waterways (e.g., narrow channels, canals, rivers and estuaries), harbour approach- and inner harbour sailing. Each of the phases has different demands and navigational resource requirements (see Table 1 for examples of different types of navigation). In open-ocean sailing; for instance, one mostly uses deduced (dead) reckoning, satellite navigation, and the radar primarily for collision avoidance. While closer to the coast, the navigators must optimise the use of all navigational tools at their disposal. Furthermore, during complex inland waterways and harbour navigation phases, an additional navigation resource is supplied for the bridge: the maritime pilot.

Table 1: Types of navigation (Hareide & Ostnes, 2017).

Navigation tool	Explanation
Deduced reckoning	Deducing position by advancing a known position for course and distances.
Satellite navigation	Use radio signals from satellites to determine the position, known as the Global Navigation Satellite System (GNSS).
Radar navigation	Using electromagnetic waves to determine the distance from or bearing to an object (different than radar for collision avoidance).
Radio navigation	Determine position using radio waves.
Piloting	Using experienced seafarers with local expertise where there is a need for continuous determination of position.

2.2 Maritime pilotage

The use of a qualified local navigator for ships in- and outbound of ports has been widely used throughout the shipping history. The early "pilots" from the ancient Greek and Roman times were mostly local fishermen who were employed by incoming ships to aid navigation. Piloting was formalised by local governments already in 1850; however, it was formally recognised by IMO in 1968 in the "Assembly Resolution A.159 (ES.IV) Recommendation on Pilotage". Today, pilotage can be defined as:

> *"The navigation and steering of a ship in a sea area, in which task comprehensive knowledge and experience concerning the local conditions of the sea are necessary for safe, economic and environmentally acceptable sailing of a ship to the desired target."*
> *– Norros (2004, p. 184).*

Regardless and important to emphasise: the captain remains responsible and accountable for the safe navigation of the ship. This responsibility and accountability entail that the master of the ship can relieve the pilot of service and request for a different pilot if the first pilot jeopardises the ship's safety. While pilots exist to increase the safety and efficiency of the operation, accidents with pilots onboard still occurs. The frequency of accidents is low; however, the consequences of such accidents can be devastating.

Accident reports show that more work is required to prevent and mitigate unfortunate accidents involving pilotage operations, such as the Godafoss, Federal Kivalina and Crete Cement accidents (Accident Investigation Board, 2010a, 2010b, 2012). The recent KNM Helge Ingstad accident also involved a pilot; however, mind that the current report is only preliminary and not conclusive (Accident Investigation Board, 2018). Accidents like these are costly and trigger attention and research to the matter of pilotage competencies.

In fact, Chambers and Main (2015) found that maritime pilots experience various physical, behavioural, and cognitive fatigue symptoms during their shifts. These symptoms can often aggravate the complexity of pilotage operations and, coalesced with the recent accident reports, stresses the importance of developing measures for carrying out safe and efficient pilotage operations.

Figure 1 below shows two pilotage signals for ships indicating the need for a pilot (left) and that a pilot is onboard (right). These signals exemplify a safety measure in pilotage operations.

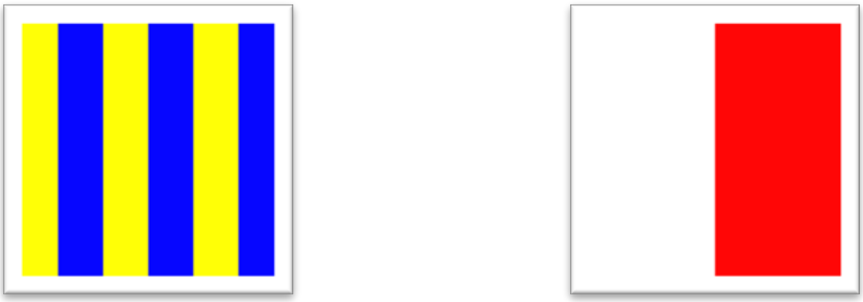


Figure 1: Signal flags in relation to pilotage operations. To the left, signal flag G indicating that a vessel requires pilot. To the right, signal flag H indicating that they have a pilot onboard.

2.2.1 The pilot-bridge team

A bridge team is typically composed of a helmsman, a navigation officer, and a captain. There can also be additional roles depending on the ship and situational requirements, e.g., a lookout and extra navigation officers. Also, if the ship is approaching a port that requires the use of pilot services; a pilot must be transported to the ship (either via helicopter or via a more frequently used pilot boat). It is worth emphasising that this requires a functional knowledge integration of two high-ranking and experienced professionals with individual backgrounds. This interplay can confound the teamwork environment (Norros, 2004).

Regardless of the team composition, however, effective teamwork must be exercised to ensure that the bridge team and the pilot safely and efficiently voyage the ship to and from the harbour. In fact, a 2010 STCW amendment was added to the list of minimum competency requirements for the captain detailed in Table A-III/II/I of the 2010 Manila amendments to the STCW Convention and Code (IMO, 2011). It accentuates and stresses the role of teamwork. The applied terminology for teamwork on the bridge is placed under the *bridge resource management* (BRM) umbrella, which was initially adopted from aviation (named crew resource management). However, “teamwork” is used throughout the doctoral research.

Moreover, the ship bridge is often designed for dedicated maritime operations, e.g., dynamic positioning and seismic geolocating services. There are also environmental conditions that factor in when configuring the staffing of the ship’s bridge, which not all are relevant to the research presented in this dissertation. The description of the different roles on the bridge is therefore restrained to the pilot, captain, navigator, helmsman, and lookout. When the pilot is integrated with the bridge team, the team is in this dissertation then referred to as the “pilot-bridge team”.

2.2.1.1 The pilot

The maritime pilot is a local navigational expert transported onboard the ship to partake the bridge team’s port approach. Pilots are (generally) experienced seafarers and differ from the onboard bridge team members as he or she is solely a transient member. The pilot, when transported onboard the ship, is then integrated with the onboard bridge team to assist in the navigation and manoeuvring of the ship. The pilot knows the fairway, seabed, buoys, quays, currents, tide and planned work well. His or her training and competency requirements are formalised in the IMO Resolution A.960(23). Other responsibilities for the pilot consist of having knowledge about safe standard routes, while being able to adapt the route according to specific ship and time constraints, hydro-meteorological conditions, and traffic. They also ensure that the ship's transit respects and conserves local interests, such as minimal pollution and noise. The pilot also knows the local services that are available and can act as an intermediary to these services when needed. Due to being part of the local organisation and services, pilots tend to have a strong but informal working relationship with the vessel traffic services (VTS), facilitating the ship crew's external communication and coordination when the ship is in- or outbound. Pilots, then, have a critical role in ensuring the safe navigation of ships in their care and is included as a transient member of the bridge team during pilotage operations.

2.2.1.2 The captain

The captain is also referred to as the skipper or master. The word "captain" is believed to derive from "caput", the classic Latin word for head. It may have been combined with "katepano", which was used as a title for a senior Byzantine military rank. Today, a maritime captain for the largest vessels is a high-grade licensed mariner: a master mariner. The minimum requirements for such a licence are regulated in the STCW Code Section A-II/1 – chapter 2 (IMO, 2011), stating that competencies in administrative and operational responsibilities must be demonstrated.

Administrative responsibilities for the captain are ensuring compliance with all laws and regulations that apply to the ship at the national, local- and international level. He or she is also responsible for following the company's procedures and policies. The list of administrative responsibilities is increasing, rendering their entitlement increasingly complex. For instance, the captain must now contend with new personnel, signing documents, unions' work-hour rules, and to further the logging and recording documentation to name a few. Regardless, the captain must still strictly ensure the safety and efficiency of the ship's operational position and has the ultimate accountability of the voyage.

The operational requirements for the captain include taking responsibility for safe navigation of the vessel. This responsibility cannot be superseded by onshore advice, the Coast Guard, or surveyor agents. In this requirement is ensuring clean and safe crew accommodation and public areas, cargo delivery, and the ship's seaworthiness. However, recent and prospective technological developments may change the way captains must ensure safe navigation of the ship (Porathe, 2019). Some changes are already happening; many captains are already experiencing reduced crew onboard, increased automation and communication technology, and new requirements like 24-hour communication accessibility of the vessel, which could further increase stress (Sellberg & Susi, 2014). Prospective changes like increased autonomy and technologies which enables remote operations and assistance may further complicate the captain's quest for safe navigation. Regardless of the responsibilities, he or she is encouraged to closely collaborate and take benefit of the other highly trained members on the bridge.

2.2.1.3 Other central roles on the bridge

The helmsman (or helm) steers the ship. He or she receives orders from the captain or the officer of the watch in the absence of the captain. Typical commands for the helm are rudder commands (i.e., a single-event change of rudder angle) and heading commands (i.e., continuous actions required to maintain a specified heading). The helmsman is expected to close-loop orders from the captain to ensure a mutual understanding and that the message is correctly received and interpreted. Steering a large ship is complex as the helmsman must understand the ship-handling of the ship following hydrodynamic forces such as ship-passing, shallow-water effects, and ocean currents.

The helm takes much help from the "lookout" who has the responsibility of observing the surrounding landscape for other ships or hazards. He or she timely provides this information to the rest of the bridge team. Depending on the crew size and the complexity of the operation, the responsibilities of the lookout could be re-assigned to other functions also.

2.2.2 Breaking down pilotage operations

Norros (2004) did a comprehensive work of breaking down a generic pilotage exercise into navigation and steering tasks. The tasks are organised hierarchically and sequentially as seen in Figure 2 below. It positions the pilot's expertise as an integral and necessary part of safe sailing. The organisation is generic and can be used as a framework in different types of pilotage, both when sailing close to port and when sailing through inland waterways.

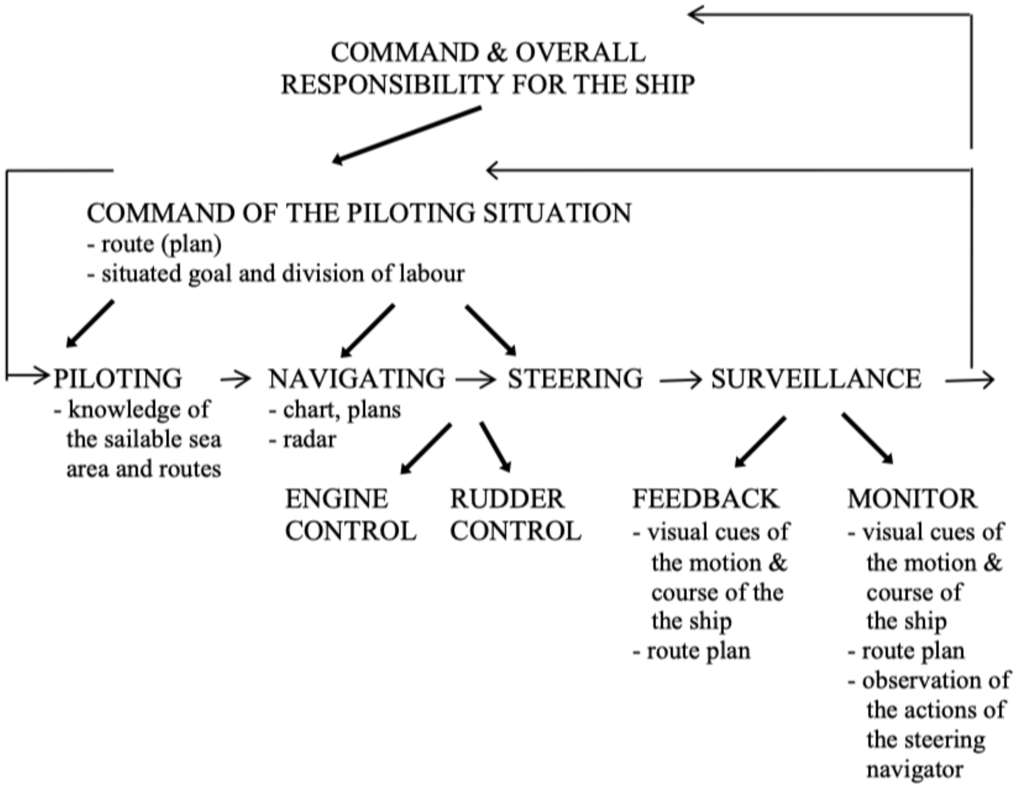

COMMAND & OVERALL RESPONSIBILITY FOR THE SHIP
COMMAND OF THE PILOTING SITUATION
- route (plan)
- situated goal and division of labour
PILOTING
- knowledge of the sailable sea area and routes
NAVIGATING
- chart, plans
- radar
STEERING
SURVEILLANCE
ENGINE CONTROL
RUDDER CONTROL
FEEDBACK
- visual cues of the motion & course of the the ship
- route plan
MONITOR
- visual cues of the motion & course of the ship
- route plan
- observation of the actions of the steering navigator

Figure 2: Hierarchical and sequential structure of the navigation and steering task in piloting situations (Norros, 2004).

Pilotage can be distinguished between piloting ships in the proximity of ports and by piloting ships through inland waterways and archipelagos (although there are regional differences). The same literature describes four methods for port pilotage and three methods for sea pilotage that are relevant to Finnish sail routes, which can also provide useful distinctions and nomenclature for pilotage operations in neighbouring countries.

However, this distinction is not absolute considering the difficulty of categorising piloting operations in general as they are all carried out in unique, open and dynamic environments.

For port pilotage, the four methods use manual control of the ship (i.e., not sailing using the autopilot), but vary in the organisation of situational command and responsibility of manoeuvring, e.g., captain is berthing the ship. For sea pilotage, one method is where the pilot is in charge of the situation and navigates in assistance of the helmsman with the captain monitoring the operation. Another method is where the pilot is in charge, but steers using autopilot while the captain is monitoring. The last method is having the captain in charge and using autopilot, whereas the pilot is monitoring the operation (Norros, 2004). Please see Table 2 below for an overview of the described bridge control configurations.

Table 2: Types of pilotage operations as described in Norros (2004). Port pilotage differs in situational command and is not specified further than the main distinctions and is therefore intentionally left blank.

Types of pilotage operations	Characteristics
Port pilotage A	Pilot in charge, helmsman aids in steering and berthing, captain is monitoring.
Port pilotage B	Pilot in charge, captain aids in steering and berthing.
Port pilotage C	Captain in charge, helmsman aids in steering and berthing, pilot is monitoring.
Port pilotage D	Captain in charge, pilot aids in steering and berthing.
Sea pilotage A	Pilot is in charge, helmsman aids the steering, and the captain is monitoring.
Sea pilotage B	Pilot is in charge, autopilot is used for steering, and the captain is monitoring.
Sea pilotage C	Captain is in charge, autopilot is used for steering, and the pilot is monitoring.

There are advantages and disadvantages for each of the methods. For instance, pilots are generally well acquainted and accustomed with sailing various kinds of ships while having expert knowledge of the local fairway. However, the pilots will not have the same competency as the crew with regards to each ship's hydrodynamic peculiarities (although exceptions apply). Ultimately, the technical tasks necessary for successful navigation in piloting operations are addressed extensively in various assessment standards at various navigational simulator training facilities following the IMO model courses (Ali, 2006).

Furthermore, Norros (2004) found that the operations were mostly pilot-centred (for both port- and sea pilotage); that is, the pilot was in charge and steering the ship.

However, further studies, e.g., Lappalainen, Kunnaala, and Tapaninen (2014), as well as anecdotal experiences from both captains and pilots indicate and suggest domain- and culture dependent variations. For instance, specialised seismic vessels could be challenging for a pilot to berth, in which case the onboard crew would “take her in”. Cultural dependent variations encompass, for instance, that the captain may advocate that only the master of the ship should be the one who berths the ship (and not the pilot). Regardless of which approach is carried out, clear articulation of which approach, i.e., clarifying tasks and responsibilities, is an essential characteristic for successful piloting (Lappalainen et al., 2014); illustrating the paramount importance of effective teamwork in pilotage operations.

2.2.3 Team and teamwork in pilotage operations

A team is defined as “two or more individuals with specified roles interacting adaptively, interdependently, and dynamically towards a common and valued goal” (Dyer, 1984; Salas, 1992). There are different types of teams depending on their application. For the pilot-bridge team in pilotage operations, McIntyre and Salas’ (1995) definition of tactical decision-making teams can be considered appropriate: these are teams that may have to operate with taskwork under time-pressure, in which the error consequences are immediate and may be severe. Teamwork is critical in tactical decision-making teams.

Pilot-bridge teams could also be considered a swift starting action team (Andresen, Domsch, & Cascorbi, 2007). These teams must perform in unfamiliar team configurations, often concerning tasks that possess a risk of immediate and severe consequences (McKinney Jr, Barker, Smith, & Davis, 2004). This characteristic may have implications for the teamwork on the bridge, such as the development of shared mental models.

Dedicated research on pilotage also supports the need to examine the social aspects of these operations. One study found that in six out of seventeen investigated piloting operations, a shared mental model of the situation was lacking among the bridge team-members (Norros, 2004). This lack of a shared mental model can suggest that the team-members assume that the other members have similar mental models of the ongoing operation (Singer & Fehr, 2005). This discrepancy in their theory of mind is further aggravated by the broad international reach for most ships, which accentuates the need for multi-cultural understanding for all pilot-bridge members, irrespective of their technical competencies. Furthermore, a study on pilots from different European countries recommended further research on systems that support the human component of pilotage operations, such as communication protocols (Gruenefeld et al., 2018).

It has also been found that even if sufficient technical competency exists within the pilot-bridge team, a functioning teamwork is critical for the safety and efficiency of the operation (Lappalainen et al., 2014; Wild, 2011). In an earlier study, Norros and Hukki (2003) found that the practised pilotage method dynamically adapts to the characteristics of the operation (harbour vs sea passage). The approach used for pilotage depends on the availability of the bridge crew rather than on the available navigation and manoeuvring technology, which have implications for the cooperative behaviour on the bridge (Norros & Hukki, 2003).

Teamwork is not crisply defined (Salas, Sims, & Burke, 2005), but relates to activities serving to strengthen the quality of the team's functional interactions, relationships, cooperation, communication, and coordination of the various team members (McIntyre & Salas, 1995). This description of teamwork puts boundaries when assessing teamwork competencies as it is described as something more than mere team performance. In the example of pilotage operations: only assessing the team's output, e.g., whether the ship is successfully berthed, will be an inadequate measure of the ship crew's teamwork competency. Instead, one must refer to the team effectiveness, taking a holistic perspective both assessing the outcome and the mechanisms that ensured that the outcome was achieved.

Please see Section 3.2 for a presentation of the teamwork literature used in the current doctoral research. Next is a description of the unit of assessment and its system properties.

2.3 The unit of assessment

The maritime shipping domain is vast and dynamic. To study this domain requires careful articulation of the boundaries for the research as it can be approached from different perspectives. The following subsections aim to provide a demarcation – boundaries – for which the current doctoral research has been conducted. The boundaries are set for both the bridge team and for the system that the bridge team is operating.

2.3.1 Demarcating the pilot-bridge team's system

The pilot-bridge team studied in the current doctoral research are limited to large commercial merchant ships in pilotage operations. This delimitation includes cargo ships, bulk carriers, oil tankers, roll-on roll-off (ro-ro) ships, and cruise ships operating in connection to Scandinavian waters. Although the content of the thesis could be applied to different types of maritime and non-maritime operations and geographical areas, it was not the focus of this research.

For carrying out pilotage operations, the pilot-bridge team operates as a joint cognitive and socio-technical system. This distinction raises the need for making the system factors and boundaries explicit (Hollnagel & Woods, 2005, p. 67). These concepts will therefore be outlined and connected to the current research.

2.3.1.1 Joint cognitive systems

A joint cognitive system is capable of anti-entropic behaviour, i.e., it can adapt its behaviour to current and anticipated environmental demands. This behaviour is a necessary element for the pilot-bridge team system's performance (and should be reflected in its assessment). The pilot-bridge team comprises several cognitive sub-systems, in which at least one sub-system must be anti-entropic for it to be classified as a joint cognitive system. The (human) navigator; for instance, can adapt his or her behaviour to environmental demands through coagent dependencies with other system components, such as updating the ship's speed and heading based on feedback from the RADAR and ECDIS.

A joint cognitive system needs a holistic approach for understanding the system's functions: For instance, a critical and typical process that desires holistic advancements in joint cognitive systems is the human – technological ensemble, just mentioned with the RADAR example above. By investigating the ensemble's performance in the system, one can advance the understanding of otherwise impenetrable performance data, in contrast, to merely considering a navigator's isolated understanding of a static RADAR image (Hollnagel & Woods, 2005, pp. 67–68). The performance assessment of the pilot-bridge team, thus, corresponds to the system's operational outcome, and not the outcome of the respective individual team member (or agent).

While cognition research is interested in "what we know", it manifests differently in individual research and systems research. For individual research, the researchers are interested in the mental processes concerning what people know, and for systems research, researchers are interested in the system processes for understanding what it knows. A joint cognitive system classification of the bridge team is, therefore, helpful for capturing the system processes. To appreciate this, one must consider that a system is more than the set of its elements (i.e., the relationships among the system's elements are substantial for its identity) and that the structure of the system elements impacts and determines its function (Ropohl, 1999). The x-axis in Figure 3 below expresses neighbouring joint cognitive systems that are close to the pilot-bridge team.

The pilot-bridge team system, like any system, is subject to the principle of excluded reductionism (Ropohl, 1999). This reductionism implies that the team system performance cannot be described (and assessed) by considering the individual, the team, or the organisational level of analysis in isolation. The renown Hawthorne experiments, first investigation reported in Landsberger (1958), early alluded towards the socio-psychological finding that an individual's behaviour cannot be understood adequately if social structures are disregarded. Clearly, including all known, unknown and unknown-unknown system effects is beyond the scope of this doctoral research and a pragmatic approach is necessary. The current research, in light of this discussion, is focused at the team level of analysis. The team level is expressed by its position on the y-axis in Figure 3 below.

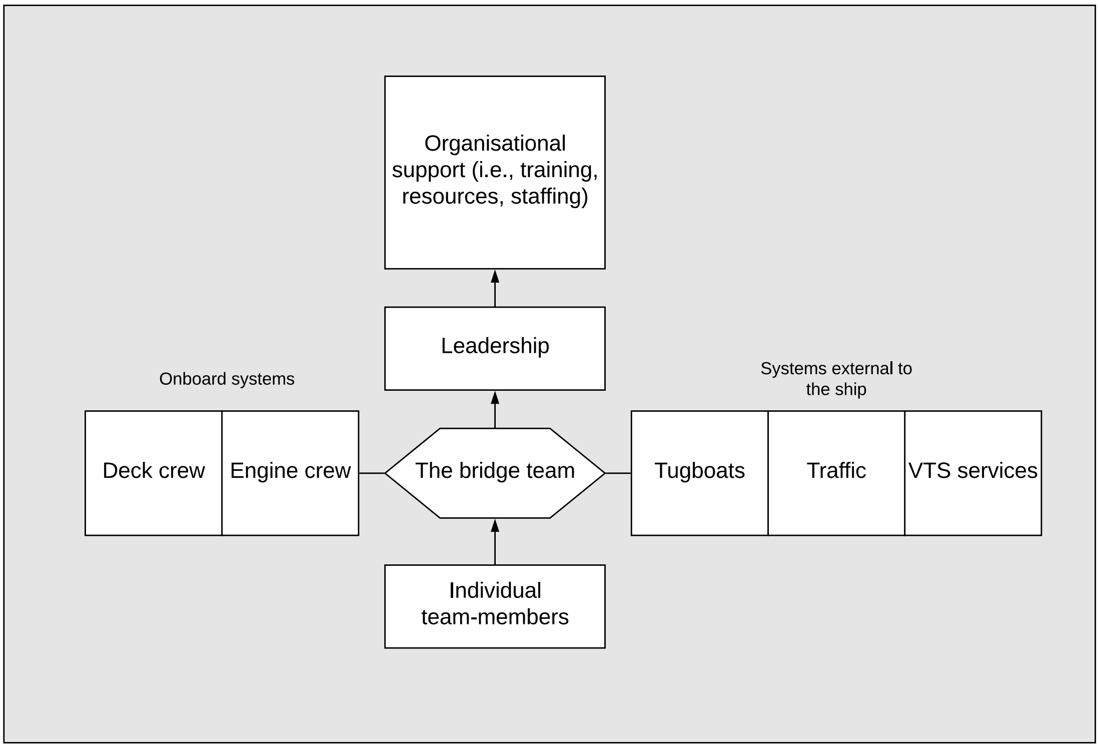


Figure 3: Visualising the bridge team's position among joint cognitive systems on the ship in conjunction to the four levels of analysis. The horizontal axis shows neighbouring joint cognitive systems, the vertical axis illustrates the classical four levels of analysis (individual – group – leadership – organisation) model.

While the bridge system is recognised as a joint cognitive system at the team level of analysis, as expressed by the two axes in Figure 3 above, four auxiliary dimensions must still be addressed to consistently delimit the properties and boundaries of the system (Hollnagel & Woods, 2005, p. 67). The researcher will therefore bring attention to (1) the system's objects, attributes, and exemplify system relationships that have been used in the doctoral research, (2) discuss external impacts on the bridge system, (3) address the complexity within the bridge system; and finally, (4) provide an overview of the bridge system by describing socio-technical characteristics applicable to the bridge team system. Defining and discussing these system dimensions are central to classify the unit of analysis properly.

2.4 The pilot-bridge team's system properties and boundaries

In Hall and Fagen (1968, p. 82), a system is defined as "*a set of objects together with relationships between the objects and between their attributes*". In this, relevant objects, attributes and relationships for the bridge team and the context in which it exists are defined to articulate the unit of analysis aptly. Subsequently, a discussion of complexity according to Flach (2012), whereas Ropohl (1999) and Vicente (1999) are used to examine the socio-technical characteristics of the bridge team.

2.4.1 Objects, attributes and relationships in the joint cognitive system

Objects are physical or abstract elements of the system, and their attributes point to the object's properties (Hall & Fagen, 1968). Objects are the components of the system, such as a button on a piece of machinery. The attribute, then, is the object's property, such as the state of the button (e.g., on/off). While an object can have a multitude of both primary and secondary attributes, only the primary properties of relevant objects are used to classify the bridge team system. In light of this, the objects and attributes that were used in the doctoral research are listed in Table 3 below. Moreover, as previously affirmed, the maritime domain is vast and diverse: This diversity is also reflected in the variations of objects found across different ship bridges. Considering this variation; a third and fourth column in Table 3 below is dedicated to highlighting system properties that diverge across maritime operations.

Table 3: List of objects and main attributes in the maritime scenario designed for the experiment reported in Article 4. MFD = Multifunctional display (i.e., can be set for different attributes).

Objects	Main attributes	Customisable object	Customisable attribute
Monitor 1 (MFD)	ECDIS, GPS overlay, AIS	Yes	Yes
Monitor 2 (MFD)	RADAR, GPS overlay, ARPA	Yes	Yes
Monitor 3 (MFD)	Binocular view	Yes	Yes
Monitor 4 (MFD)	Conning: speed and heading	Yes	Yes
Monitor 5 (MFD)	Conning: Engine throttle	Yes	Yes
Instrument 1	Throttle control	No	Yes
Instrument 2	Thruster control	Yes	Yes
Instrument 3	Rudder control	Yes	Yes
Windows	Optical view of surround area	No	Yes
Book	Logging	No	Yes
Documentation	Pilot cards	No	Yes

Radio	VHF	No	Yes
Handheld radios	UHF	No	Yes
Human 1	Captain/Officer of the Watch	No	Yes
Human 2	Pilot	No	Yes
Human 3	Helmsman	Yes	Yes
Human 4	Lookout	Yes	Yes
Human 5	Navigator	Yes	Yes
Human 6	VTS	No	Yes
Human 7	Tugs	Yes	Yes
Human 8	Other Vessels	Yes	Yes
Human 9	Skandia Harbour	Yes	Yes
Human 10	Boatswain	No	Yes

The next aspect to consider concerns the relationship between objects, which can be either direct or indirect. A direct relationship describes objects where the outcome in isolation influences the outcome of another object, and an indirect relationship corresponds to objects moderating the connection between the objects. Furthermore, seeing that objects in the system and their interdependent relationship are a combination of technical and social, the bridge team system is also socio-technical system (Ropohl, 1999). In Table 4 below, examples of simple relationships for a few objects that were used are presented: Please note that the list is only a limited example and far from exhaustive, as there often are numerous visible and subtle inter-dependent interactions in such complex operations.

Table 4: Example of system relationships by listing objects and attributes and an outcome from their relationship.

Object (attribute 1)	Object (attribute 2)	Outcome
Human (navigator)	Monitor 1 (ECDIS)	GPS position update
Instrument 1 (throttle)	Instrument 2 (thruster)	Course alteration
Coordination (Backup behaviour)	Instrument 3 (rudder)	Rectify ship heading
Schemata (job)	Human (lookout)	Knowing which external triggers to bring into team's attention

2.4.2 External impacts on pilotage operations

The environment around the pilotage operation is typically a process in constant change (Norros, 2004, p. 32), and these external environmental factors are important to consider for open systems (Flach, 2012). These factors influence the system behaviours but are not part of the system itself. This external impact contrasts to closed systems where the boundary between the system and the environment is impermeable. The bridge team is an open system and has to accommodate changes in response to environmental demands continuously. For instance, what may appear as an uneventful voyage into port, could abruptly elevate due to a sudden change in the weather condition. Understanding the environmental demands is critical for the safety of the operation and as a consequence, are essential for understanding pilotage operations. For the current research, three predominating external factors are considered: weather, time pressure and traffic.

Weather and hydrodynamic forces impact the ship's manoeuvrability and navigation. While this impact is less problematic in open ocean (relative), it can significantly impact the complexity of the operation in shallow water and during berthing, e.g., depth effects (Wang et al., 2017). Please see Figure 4a and 4b below for an illustration of six ship motions that can be influenced by hydrodynamic forces and the manoeuvrability of the ship.

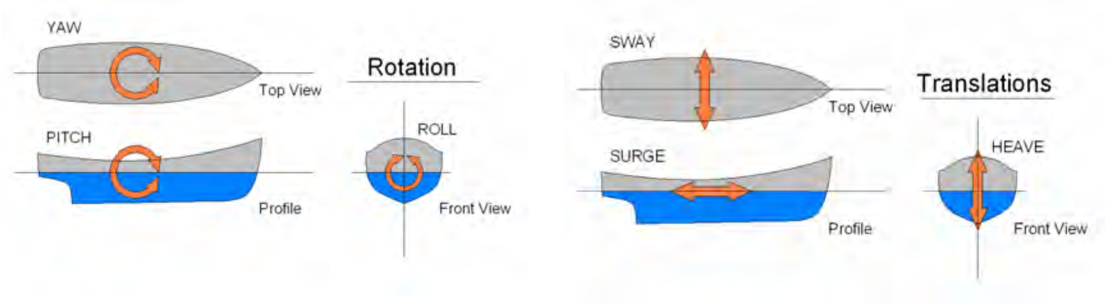


Figure 4a and 4b: Illustrations of the six ship motions. Figure A shows the three rotational degrees of freedom whereas Figure B shows the three translation degrees of freedom. Both figures are public domain, created by Wikipedia user: Jmvolc.

Time is another external dimension that impacts the (perceived) pressure of the operation. Different elements contribute to time pressure; for instance, weather-invoked time pressure, tidal dynamics (certain areas are only passable during high-tide), pressures coming from the port (schedules), and ship-owner pressure (e.g., economic interests). Regardless of the source, time-pressured situations are found to negatively impact the operator's decision making (e.g., Wickens, Stokes, Barnett and Hyman's (1993) study on aviation pilots in time-pressured situations). There are several prevailing theories attempting to explain how time-pressured situations impact skill performance:

two of them are self-focus and explicit monitoring (Baumeister, 1984). The line of reasoning is that performance pressure increases anxiety and self-consciousness about performing correctly; which subsequently activates an intrinsic and cognitive step-by-step-control for performing the necessary set of actions, in which the mechanism is further explained in the renown skill-, rules,- and knowledge model described in Rasmussen (1983). This step-by-step control inhibits the automated skills to operate as already developed through hours of experience and training. A consequence, in turn, is excessive use of working memory, reducing the operator's capacity to, for instance, project future actions (Baumeister, 1984; Endsley, 1995).

On the contrary, there are certain time-pressured conditions found to impact performance positively. For instance, conditions that are repetitive and too low in demand can pacify operators, reducing their operational vigilance (Endsley, 2017), suggesting that some level of time pressure is necessary. Identifying and delimiting time pressure is an important aspect that impacts the performance of complex joint cognitive systems.

Maritime traffic is the third external factor that will be addressed. Traffic density has the potential of making navigation – especially in constrained waterways – excessively complex. The surrounding vessels operate, navigate and resolve conflicts locally. While there exist central information services (i.e., the vessel traffic service), the management of maritime traffic is distributed (van Westrenen & Praetorius, 2014). As the world's fleet capacity and need for shipping increases (UNCTAD, 2018), the traffic complexity will naturally also increase. There are studies suggesting remedies for the increase in traffic complexity, e.g., by organising a centralised planning and coordination system (van Westrenen & Praetorius, 2014). However, the maritime industry is slow to change, and in the interim, traffic complexity maintains an external dimension that must be carefully monitored by the pilot-bridge team.

Collectively, weather, time, and traffic factors contribute to raising the complexity of the joint cognitive (bridge team) system in carrying out a safe and efficient voyage. With regards to the internal and external system factors, maritime navigation operations are considered complex as defined in Flach (2012).

2.4.3 Complexity in pilotage operations

A pilotage operation is considered as high-risk and requires that the pilots perform complex procedures in unfamiliar team configurations (Andresen et al., 2007; Darbra, Crawford, Haley, & Morrison, 2007). For instance, the operation has uncertain hydrodynamic processes and a high degree of complexity (in this example: a technology-mediated information representation), as suggested in Norros (2004).

Complexity is a function of the number of objects in the system (dimensionality) and the relationship between these objects (Flach, 2012), where more objects and a higher number of interdependent relationships are considered more complex. Complexity, in turn, increases with the addition of interdependent factors (i.e., increased interdependent dimensionality). For maritime pilotage operations, relationships between system factors are often interdependent and multidimensional, which supports the classification of maritime pilotage operations as complex. In these operations, all of the examples are interconnected, either directly or indirectly. Four simple examples of intricate relationships between critical factors are given in Table 4 above. To illustrate a simple example of navigation complexity, Figure 5 below shows a potential connection between the internal bridge system and external system that are important for the performance of the bridge team. The complexity increases when the number of dimensions (list on the left) increases and when connecting more lines in the figure to the right.

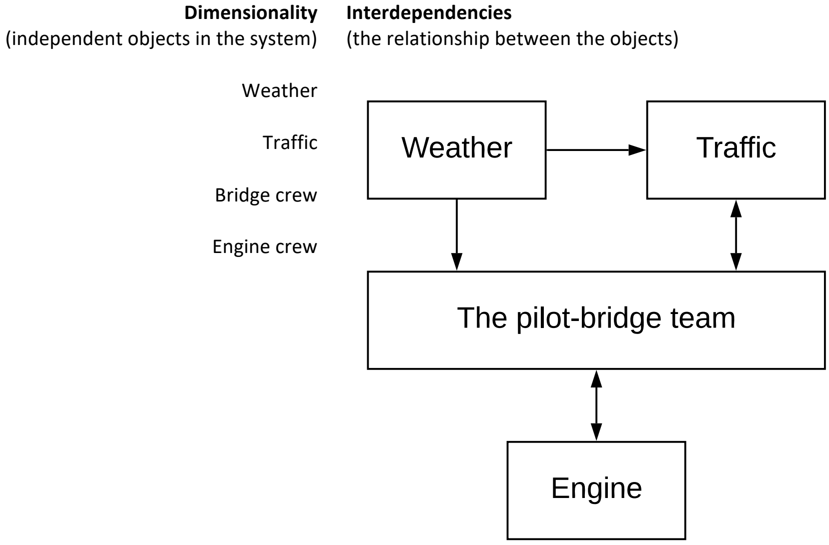


Figure 5: Simple illustration of complexity applied to navigation, retrieved from: Ernstsen, Musharraf, Mallam, Nazir and Veitch (2018).

2.4.4 Socio-technical characteristics in pilotage operations

Analyses of complex socio-technical operations have been carried across a multitude of disciplines, like aviation (de Carvalho, Gomes, Huber, & Vidal, 2009), healthcare (Carayon & Buckle, 2010), military (Rafferty, Stanton, & Walker, 2010), cars and railroad (Stanton & Salmon, 2011), as well as for maritime operations (Stanton & Bessell, 2014). Vicente (1999) formalised a set of complex socio-technical characteristics that have subsequently been adapted to maritime settings; for instance, Praetorius and Hollnagel (2014) applied these characteristics to vessel traffic services, and Ernstsen et al., (2018) described the characteristics in relation to pilotage operations (see Table 5 below). While the list is not covering all aspects, it does give an overview of complex socio-technical characteristics that can exist in various maritime pilotage operations.

Theories on joint cognitive systems (Hollnagel & Woods, 2005), research on operational complexity (Flach, 2012), and socio-technical system frameworks (Ropohl, 1999) are employed for defining the pilot-bridge team system and to provide background and context for the assessment tool.

The next chapter will be addressing training- and assessment theories relevant for the development of the tool.

Table 5: Socio-technical characteristics applied to pilotage operations, retrieved from Ernstsen et al., (2018).

Characteristic	Maritime application
Dynamic	A high lag between action and response: workers must anticipate vessel's future state.
Large problem space	The complexity of work based on a plethora of relevant variables (ship technical, human-related and environmental).
Social	Bridge system functioning is dependent on collaboration and cooperation.
Distributed	Dependent personnel involved in complex navigations are culturally and geographically distributed (e.g., engine room crew, tugboat operators, port authorities).
Heterogeneous perspectives	Operators with different backgrounds and potentially conflicting values are common in the maritime industry.
Hazard	There is a high degree of a potential hazard upon failure.
Couplings	Bridge system depends on complex subsystems (e.g., engine room, VTS, surrounding vessels), which makes it difficult to predict all effects of an action.
Automation	Computer algorithms control work operations while bridge personnel often monitor, thus operators are rendered unaccustomed to perform compensatory activities upon system deviations.
Mediated interaction	Bridge operators must rely on interfaces to acquire a representation of the system state.
Uncertainty	Sensors and indicators monitoring the technical system of the vessel may provide the operators with erroneous information.
Disturbances	The bridge system is dealing with unanticipated events which require improvised action (e.g., when checklists and standard operating procedures are mis-fitting) to rectify system deviations, thus requiring that operators possess a conceptual understanding of their work.

3 Theoretical framework

Training and assessment are two significant fields of research pertinent to the development of a performance assessment measure in full-scale training simulators. This chapter, therefore, provides a comprehensive overview of the theoretical framework that is employed in the current doctoral research and describes relevant training- and assessment research from a human factors-perspective. Then the last part of this chapter expands on relevant theoretical background for teamwork.

3.1 Training and assessment

Training and assessment are indispensable elements of Human Factors (HF). The HF discipline is scientific, theoretical, and applied while dealing with psychological, physical, and organisational aspects of the interaction between humans and systems (Horberry, Grech, & Koester, 2008; IEA, 2019). It involves the study of factors and development of tools for attaining three goals: enhancing performance, increasing safety, and increasing user satisfaction (Wickens, Gordon, Liu, & Lee, 1998). While one approach may attempt to increase production output (performance) on the cost of quality (safety), a human factors' approach would aim to satisfy all goals at once (Alexander, 2002; Hendrick, 1996). For training and assessment, this requires methods that prepare the operator for the challenges he or she will encounter in their working endeavours. This requirement entails teaching, practising, and assessing the physical and mental skills demanded by the job.

3.1.1 Training

Improving the competencies of a human operator - the trainee - requires meticulous planning. There are plenty of training methods at disposal, ranging from on-the-job training, to traditional classroom training, and full-scale simulator training. However, a conundrum for any training program is to choose and employ methods that are best suited for reaching the present-day training needs (Flin & O'Connor, 2017). One critical aspect of this conundrum is to recognise the type of competency that is required for a particular task: declarative, procedural, or skills related to automaticity (Fitts & Posner, 1967; Kluge, 2014).

3.1.1.1 Declarative, procedural and skill knowledge

Declarative knowledge involves knowledge about a task, which is essential to have in any complex operation (Kraiger, Ford, & Salas, 1993). This need translates into knowing various facts about the job (and its environment); such as task information, semantics and procedures, as well as generating various relevant personal experiences tied to carrying out the task (Squire, 1992). This type of knowledge is usually not cognitively well organised; hence, it is often inefficient when carrying out a particular task (Wickens et al., 1998). Still, this knowledge is imperative in complex operations when novel situations arise that require creative problem solving (Rasmussen, 1983). Training operators to have proper knowledge about a task and the environment is critical for ensuring that the task is effectively, efficiently, and safely carried out during complex operations (Anderson, 1996).

Procedural knowledge is knowing how a task is carried out. As the operator keeps rehearsing and practising a specific task, he or she develops procedures that make it easier to carry out tasks. These procedures can be internalised and swiftly retrieved and employed when needed (Ackerman, 2014; Kraiger et al., 1993). The procedures consist of rules and if-then statements that the operator follows once the stimuli for the action is there. The complexity of the tasks put different demands on the amount of practice required before enough knowledge of how a task is carried out exist.

Finally, as the operator further cultivates his or her expertise, execution of the task might mature into automaticity and skills. This development allows the operator to carry out a task with the use of minimal cognitive resources. For complex operations during high situational demands, being able to carry out various tasks with minimal cognitive strain is advantageous for maintaining an awareness of the situation. An expert operator will be able to carry out tasks using automated behaviours, knowing when to apply the various rules and if-then procedures, and possess comprehensive knowledge of the system and its surrounding environment (Kluge, 2014). This competency makes experienced personnel invaluable assets for safely and efficiently carrying out complex operations.

3.1.1.2 Simulator training in complex operations

Cultivating the operator's expertise in complex operations; however, requires advanced training methods, considering that these operations entail and require declarative, procedural, and skill-based competencies (Nazir, Øvergård, & Yang, 2015; Nazir, Sorensen, Øvergård, & Manca, 2015; Rasmussen, 1983). Full-scale simulator training is a prevalent and essential advanced training method for improving competencies in complex operations, especially compared to on-the-job training, in which the risk (and cost) would be significantly higher (Vederhus et al., 2018).

Simulator training has been suggested as effective measures for improving operator competencies across a number of high-risk industries, such as aviation (De Winter, Dodou, & Mulder, 2012), nuclear process plants (Nazir, Øvergård, et al., 2015), as well as for offshore , railroad, and maritime (Håvold, Nistad, Skiri, & Ødegård, 2015; Nazir, Øvergård, et al., 2015) for effectively training the operators. Furthermore, for the maritime industry in particular, the numerous conventions and codes that the IMO – including the International Convention for the Prevention of Pollution from Ships (MARPOL), International Convention for the Safety of Life at Sea (SOLAS), International Ship and Port Facility Security Code (ISPS-code) and the STCW request substantial performance demands on the crew, which together emphasise a need for sophisticated full-scale simulator training (Baldauf, Dalaklis, & Kataria, 2016).

3.1.1.3 Training transfer in full-scale simulator training

A full-scale simulator allows the trainee to rehearse complicated situations in immersed settings that may arise during actual operation, and thus be better prepared to perform when needed. Immersivity is often connected to improved training transfer, which is an argument for using full-scale simulators in high-stake operations. It can also justify the extensive training cost associated with full-scale simulator training if rightly designed

(Dahlstrom, Dekker, Van Winsen, & Nyce, 2009; Dede, 2009). Besides, training simulators offer controlled environments that can be designed for the students' level of understanding and can be flexible in terms of pausing scenarios for feedback and discussion, if needed (Maran & Glavin, 2003).

At the same time, attempts to investigate the benefits of using emerging simulator systems, such as virtual reality systems, have been made, as reported in Mallam, et al., (2019), Renganayagalu, Mallam, Nazir, Ernstsen and Haavardthun (to be published) and Ernstsen, Mallam & Nazir (to be published).

For the training to have value, the acquired competencies must also be transferred from the training environment to its real-world application. Training transfer is conventionally defined as the application, generalisability, and maintenance of acquired knowledge and skills (Ford & Weissbein, 1997). A large meta-analysis and literature review summarised the past decades of transfer research and identified that Baldwin and Ford's (1988) portrayal of training transfer was still praised (Blume, Ford, Baldwin, & Huang, 2010). This model of the transfer process (Figure 6) organises training inputs, training outputs and transfer conditions as three critical variables for successful training transfer, and the criteria in large parts are relevant in today's contexts as well (B. S. Bell, Tannenbaum, Ford, Noe, & Kraiger, 2017).

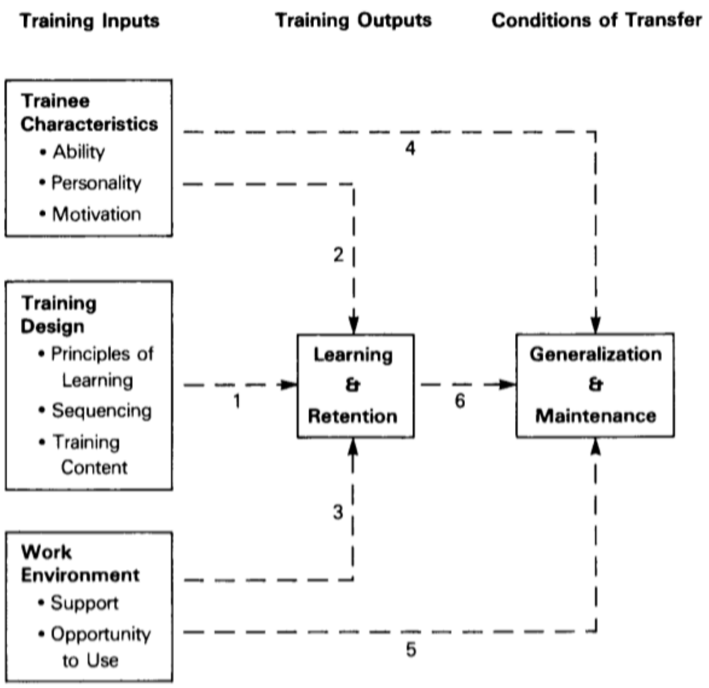


Figure 6: Model of the training transfer process (Baldwin & Ford, 1988).

3.1.1.4 The role of performance assessment

Effective performance assessment is essential for safe and efficient operations. The necessity for valid and reliable assessment is multi-dimensional: the assessment is important for 1) the individual trainee (Sadler, 1989), 2) the stakeholders; for instance, the company that hires the trainee (Taras, 2005), 3) the trainer and training program design (Taras, 2005), and 4) a company's competency modelling and job-needs analysis (Ruggeberg, 2007).

Strategic decisions across these dimensions are regular; whether it is to hire (or not hire) an operator, to further (or not) refine skills of an operator, or to reshape the training and educational program. Increasing the precision and consistency of the performance assessment methods is a prerequisite for effective strategic decision making, as it will provide the decision makers with more accurate information about the competency needs of the workforce (Bowen, Ledford Jr, & Nathan, 1991; P. J. Taylor, Driscoll, & Binning, 1998). This decision making, then, has subsequent implications for operational performance. The next section expands on the performance assessment literature.

3.1.2 Performance assessment

Assessment refers to a judgement justified according to criteria, their weightings, and the selection of specific goals (Scriven, 1967, p. 40); however, the specific nomenclature for the term "assessment" varies across disciplines and languages, where "assessment" and "evaluation" are commonly interchanged. For precision, the doctoral researcher intends to use assessment as here defined and consistent with the reasoning in Taras (2005). The term evaluation refers in this dissertation to judgements about courses and course delivery in educational settings.

3.1.2.1 Formative and summative assessment

There exists a plethora of approaches and perspectives for achieving the correct assessment of performance depending on its purpose. Two important distinctions are formative and summative assessment, as introduced in Scriven (1967). Formative assessment entails basing a performance assessment on multiple criteria and with a purpose of further shaping and improving a trainee's competency. In contrast to summative assessment, which aims to conclude and summarise the trainee's performance (Sadler, 1989).

Formative assessment is essential in successful education programs. A key aspect of formative assessment is feedback. One conceptualisation of feedback recognises it as information given by a trainer (or other agents) regarding aspects of the trainee's performance (Hattie & Timperley, 2007), feedback is thus a consequence of a trainee's performance and a central part of formative assessment (Taras, 2005). Moreover, feedback must provide information that closes the gap between what is understood and what is aimed to be understood (Ramaprasad, 1983; Sadler, 1989). This necessity is well-put in Ramaprasad (1983): "information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way" (p.4).

The other distinction, summative assessment, encapsulates all evidence up to a given point and summarises it for the trainee, e.g., as a grade score. This approach contrasts formative assessment as the feedback in the summative approach is not designed for exposing which topics the trainee needs to improve. Where formative assessments rely on explicit formulations of criteria, a summative assessment can rely on either explicit or implicit criteria.

These implicit and explicit criteria are manifestations of what is considered essential and relevant for any judgment in any context. An implicit criterion resides in the head of the evaluator, which subtly and meaningfully impacts the judgment. An explicit parameter, however, is formulated and can, therefore, more easily be shared between the assessors (Taras, 2005). This distinction of implicit and explicit assessment criteria has implications for the human bias in performance assessment.

3.1.2.2 Human bias in performance assessment

Human assessors are continuously prone to different types of assessment bias and must therefore rely on these explicit parameters to improve their reliability (Moorthy, Munz, Sarker, & Darzi, 2003). Examples of biases that impacts assessment are serial positioning effects, whereby one has a better memory of initial and final actions (Murdock Jr, 1962); halo effects, the cognitive bias whereby the initial – and often ambiguous – assessment of a trainee impacts on his or her subsequent assessment (Nisbett & Wilson, 1977b); and recognition-primed inferences, whereby one favours actions that are familiar to the assessor – typically crediting trainees' ability to solve tasks on the basis of how the assessor would accomplish it (T. D. Wilson & Brekke, 1994). These biases significantly affect what information is retrieved and evaluated (Nisbett & Wilson, 1977a).

Tversky and Kahneman (1974) presents three heuristics (i.e., cognitive shortcuts) that people make when making judgements under uncertainty: representativeness, availability, and anchoring. While these heuristics are economical and generally effective, they also lead to systematic and predictable errors. When humans are assessing operators in full-scale simulators, they are likely prone to the same heuristics.

Representativeness is defined as "the degree to which an event is similar in essential characteristics to its parent population and reflects the salient features of the process by which it is generated" (Tversky & Kahneman, 1974). This heuristic entails that people use categories when making inferences and assessments. This effect was found in a psychological experiment where students were asked to estimate the grade point average of hypothetical students. The group of students that received descriptive information about the hypothetical students ignored relevant statistics when asked to estimate the grade point average (Kahneman & Tversky, 1973). This effect has been an argument for not basing admissions on interviews and could also be considered as a pitfall in maritime full-scale simulator assessment as well.

Availability heuristic is employed when people assess the frequency of a class or the probability of an event by the ease with which instances or occurrences could be brought to mind (Tversky & Kahneman, 1974). This cognitive shortcut is useful for assessing frequencies and probabilities of a situation because it favours the larger and more frequent instances. However, there are human biases in the retrievability of the instance. In an experiment, people were asked to hear a list of well-known personalities of both sexes and were subsequently asked to judge if the list contained more men or women. In some lists, the men were more famous, and in other lists the females were more famous. Consequently, for all lists, the subjects erroneously judged that the classes consisting of the more famous personalities to also be more numerous (Tversky & Kahneman, 1973).

Furthermore, the retrievability of instances are also impacted by the salience of the particular instance (Tversky & Kahneman, 1974). An example for full-scale assessment, dominant actions carried out by the students and trainees could be more easily retrieved in the assessor's memory and subsequently be weighted unfair (in either direction) in relation to the students' and trainees' overall performance. Finally, the bias of imaginability plays an important role in the evaluation of performance and risk, in which the vivid and adventurous actions and expeditions are often more easily retrieved for the assessor (Tversky & Kahneman, 1973). This effect may lead the assessor to grossly underestimate possible dangers that are difficult to conceive (or come to mind) compared to dangers that are more colourful. The ease with which instances or occurrences are brought to mind is a significant human bias in judgement.

Anchoring is a third heuristic that Tversky and Kahneman present in their seminal paper. This heuristic represents when people make estimates by starting from an initial value which is adjusted to yield the final answer. However, regardless of the initial value's source, the adjustment for the final answer was mostly insufficient (Slovic & Lichtenstein, 1971). This effect has been (repeatedly) demonstrated by having subjects watch a number being randomly generated between 0-100, and then asked to guess whether a given quantity is larger or smaller than the random number, e.g., "Is the percentage of African countries which are members of the United Nations larger or smaller than 65 %?". The answer correlated with the arbitrary given number (Pohl & Pohl, 2004; Tversky & Kahneman, 1974). From a different perspective, people have been convinced to buy more quantities of a product by labels such as "limit 12 per customer" (Yudkowsky, 2008). A potential consequence for the event of assessment could be that assessors are biased by prior examinations: That the final evaluation of a performance is influenced by the evaluation of the prior evaluation.

The illusion of validity when making assessments further complicates the issue of getting precise and consistent performance assessments when considered in light of the above discussion. This illusion of validity gives people confidence that they have made a correct prediction with little to no regard to disconfirming factors. In fact, people express great confidence even when the assessor is aware of the factors that limit the accuracy of the prediction (Tversky & Kahneman, 1974). For instance, a person could firmly assert that a profile bio describes a librarian even if the biographical description of the profile is vague.

In sum, people's heuristics and cognitive biases are effective for fast decisions but can also make the assessor prone for erroneous performance evaluations. One counter measure, as mentioned, is to structure the assessment using explicit assessment parameters and criteria (Taras, 2005). This formulation of the parameters also enables the option to subsequently re-assess the scenario using a second evaluator. In complex operations, systemising these criteria is important for properly understanding what is being assessed (Ernstsen, Nazir, Røed, & Manca, 2016) as people can easily underestimate the probability of failure in complex system, due to effects such as anchoring (Tversky & Kahneman, 1974). Furthermore, there are various perspectives on how data can be analysed concerning an operation that is being scrutinised, where the use of computers is becoming increasingly popular (the computer is not directly prone to human bias). The computer can be a practical tool for systemising and making the criteria for assessment more explicit.

3.1.2.3 Using computers in assessment

Computers have been assisting assessment since the 1990s (Boyle & O'Hare, 2003), and increasingly sophisticated computers and software also increases the use of computers for assessment purposes (Gekara et al., 2011). Zakrzewski and Stevens (2000) points to three advantages of computerising the assessment procedures: it reduces workload pressure (particularly for multiple-choice tests), faster feedback for students and trainees, and easier data management. Additionally, it is also suggested that computer-assisted assessment facilitates more consistent and objective assessment (Roberts, Newble, Jolly, Reed, & Hampton, 2006). Another benefit is the opportunity for self-assessment, especially for formative purposes (Hodson, Saunders, & Stubbs, 2002). At the same time, setting up a computer-assisted assessment framework is costly, but given the benefits it is often regarded as cost effective in the long-tern (Gekara et al., 2011). However, Zakrzewski and Stevens (2000) and Conole and Warburton (2005) point to, among other, the need to have measurements tools that are valid, reliable and flexible as three pre-requisites for a successful assessment framework.

The use of computers in assessment has been widely debated concerning its use for summative or formative assessment, but as discussed in Gekara, Bloor and Sampson (2011), many researchers, e.g., Conole and Warburton (2005), argue that computer assisted assessment can be used in both summative and formative assessment purposes. Benefits such as automated assessment, enabled by the use of computers, are also increasingly valued, e.g., Manca, Nazir, Colombo and Kluge (2014), and it seems that considering all the benefits, that the use of computers in assessment is beneficial.

3.2 Teamwork

Teamwork can be defined as a “set of interrelated thoughts, actions, and feelings of each team member that are needed to function as a team and that combine to facilitate coordinated, adaptive performance and task objectives resulting in value-added outcomes” (Salas, Sims, & Klein, 2004). Teams are “two or more individuals with specified roles interacting adaptively, interdependently, and dynamically towards a common and valued goal” (Dyer, 1984). The teamwork requirements are expected to change depending on their application, in which tactical decision-making teams (such as teams working in complex pilotage operations) experience time-pressure and encounter error consequences that are immediate and severe (in contrast to, for instance, teams working in offices). For tactical decision-making teams, then, effective teamwork is essential.

However, the specific teamwork requirements for tactical decision-making teams are also subject to change depending on the application (Priest, Burke, Munim, & Salas, 2002). While numerous team taxonomies have been developed; for instance, Devine (2002); Marks, Mathieu and Zaccaro (2001), and Sundstrom (1999). It is argued that the team- and taskwork requirements of specific teams need to be analysed and understood in its context to augment team effectiveness (Salas et al., 2005).

For instance, Rafferty, Stanton and Walker (2010) investigated teamwork in tactical decision-making teams in complex military operations and adapted leading teamwork models for the particular case of fratricide. In their research, they conducted a comprehensive literature review (n = 80 papers) spanning 30 years of teamwork research. Four of the core teamwork factors that they identified, and which are considered relevant for the current doctoral research will be briefly presented next.

3.2.1 Communication, coordination, cooperation, and shared mental models

3.2.1.1 Communication

Communication is an explicit transfer of information between individuals and must consist of a sender and a receiver (McIntyre & Salas, 1995). There are different reasons for needing to communicate and the means of doing this; and various disciplines have extensively researched the process of communicating in teams. The current paper follows the break-down of communication as described in Wilson, Salas, Priest and Andrews (2007), which also fits to the demands of tactical decision-making teams. There are particular two necessary subfactors of communication that are highly relevant in these teams: i) Reasons for communication; that is, the content of the exchanged information, and ii) how team-members exchange this information, i.e., the phraseology. The way the team uses terminology, articulate words and sentences, following standardised communication procedures (e.g. closed-loop communication: mutual acknowledge- and verify information requests) and using effective information and communication technology.

3.2.1.2 Coordination

Coordination is important for effective team performance (Guzzo & Shea, 1992; Swezey & Salas, 1992). It is often considered as the ability of team members to act in concert without the need of explicit communication (MacMillan, Entin, & Serfaty, 2004), or even to have a distributed cognition among the team agents (Hutchins, 1995). Coordination is necessary for proper sequencing, synchronising, integrating and completion of team tasks without wasting valuable resources such as time and manpower (Cannon-Bowers, Tannenbaum, Salas, & Volpe, 1995). While coordination is interdependent on other teamwork factors (such as having shared mental models) for optimal performances, three sub-mechanisms are generally found to be important for coordination: back-up behaviour, mutual performance monitoring, and adaptability (K. A. Wilson et al., 2007).

Back-up behaviour is the ability to anticipate the needs of other team members and then to shift workload among members to achieve a balance during periods of high workload (Salas et al., 2005).

Mutual performance monitoring refers to the ability to be attentive to other team member's tasks while undertaking his or her responsibilities and providing feedback about mistakes and lapses to facilitate self-correction (McIntyre & Salas, 1995). Adaptability is the team's ability to adjust their strategies following information gathered from the environment, for instance through the use of back-up behaviour (Campion, Medsker, & Higgs, 1993). This ability involves that the team stays vigilant to identify cues and subtle changes in the surrounding environment (Salas et al., 2005).

Research on coordination has demonstrated that high-performance teams are able to rely on implicit coordination strategies in time-pressured situations for a speedier exchange of information (Entin & Serfaty, 1999; Shah & Breazeal, 2010). It should be mentioned that similar findings have been found in maritime experiments as well, e.g., Espevik, Johnsen and Eid (2011) investigated coordination and performance in co-located and distributed teams in different naval operational conditions.

Please see Table 6 below for the conceptualisations of communication and coordination.

Table 6: Conceptualisations of communication and coordination.

Communication		Coordination		
Information exchange	Phraseology	Back-up behaviour	Mutual performance monitoring	Adaptability
Refers to what information is delivered between the sender and receiver (K. A. Wilson et al., 2007).	Refers to how the information is delivered between sender and receiver (K. A. Wilson et al., 2007).	The ability to anticipate other team member's needs through accurate knowledge about their responsibilities. This includes the ability to shift workload among members to achieve balance during high periods of workload or pressure (Salas et al., 2005).	The ability to develop common understandings of the team environment and apply appropriate task strategies to monitor teammate performance (Salas et al., 2005).	The ability to adjust strategies based on information gathered from the environment through the use of back-up behaviour and reallocation of intrateam resources (Salas et al., 2005).

3.2.1.3 Cooperation

Cooperation is considered the attitudinal aspect of teamwork and is an antecedent for communication (K. A. Wilson et al., 2007). Early teamwork research already identified it as an important aspect of teamwork, e.g., Siegel and Federman (1973). There are several essential factors for cooperation; however, team orientation and mutual trust are considered most relevant within the current research framework. These two factors are commonly found to serve high importance for tactical decision-making teams' cooperation (Rafferty et al., 2010; Salas et al., 2005).

Team orientation is a complex factor that should be assessed from both an individual- and group level perspective (Eby & Dobbins, 1997). Individual components of the factors refer to a person's locus of control and perceived self-efficacy, described in Bandura (1991), which are two facets of believing that one has a potential to contribute (to the team). Moreover, on the group level perspective, the team composition is important (Mathieu, Tannenbaum, Donsbach, & Alliger, 2014): such as having teams with a heterogeneity of skill and abilities within the team (Gladstein, 1984; Jackson, 1992), while also exercising homogenous attitudes and preferences, are aspects that are linked to team orientation, subsequently critical for cooperation (Ancona & Caldwell, 1992).

The other critical mechanism for cooperation is to have mutual trust in the team. It is the shared perception that team members perform actions which are important to other team members and their joint endeavour (Salas et al., 2005). Research on cooperation as a necessary attitudinal component for effective teamwork has been summarised in recent studies (S. T. Bell, Brown, Colaneri, & Outland, 2018). This study concluded that team composition shapes the affective, behavioural and cognitive components of teamwork.

3.2.1.4 Shared mental models

Everyone is continuously and implicitly affected by his or her theory of the world and is based on own life experiences (De Villiers, 2000; Goodwin, 1994; Neisser, 1976). This cognitive mechanism follows the individual into their team interactive behaviours and is also linked to team performance (Converse, Cannon-Bowers, & Salas, 1991; Fischer, McDonnell, & Orasanu, 2007). Shared mental models have been widely explored and have throughout been given different terminology and taxonomies; however, most of the theories on the topic points to its significance for team performance.

One perspective, particularly in tactical decision-making teams, is to consider a multiple mental models-approach; mental models of technology (i.e. equipment functioning), of tasks (i.e., task contingencies and procedures), of team interactions (i.e., team roles and interaction patterns), and team mental models, as in having a schemata for other team members' knowledge, skills and abilities (Converse, Cannon-Bowers, & Salas, 1993; Mohammed & Dumville, 2001). Effective mental models are important for efficient coordination; and thus, team performance (K. A. Wilson et al., 2007; Zoogah, Noe, & Shenkar, 2015).

Please see Table 7 below for the conceptualisations of cooperation and shared mental models.

Table 7: Conceptualisations of cooperation and shared mental models.

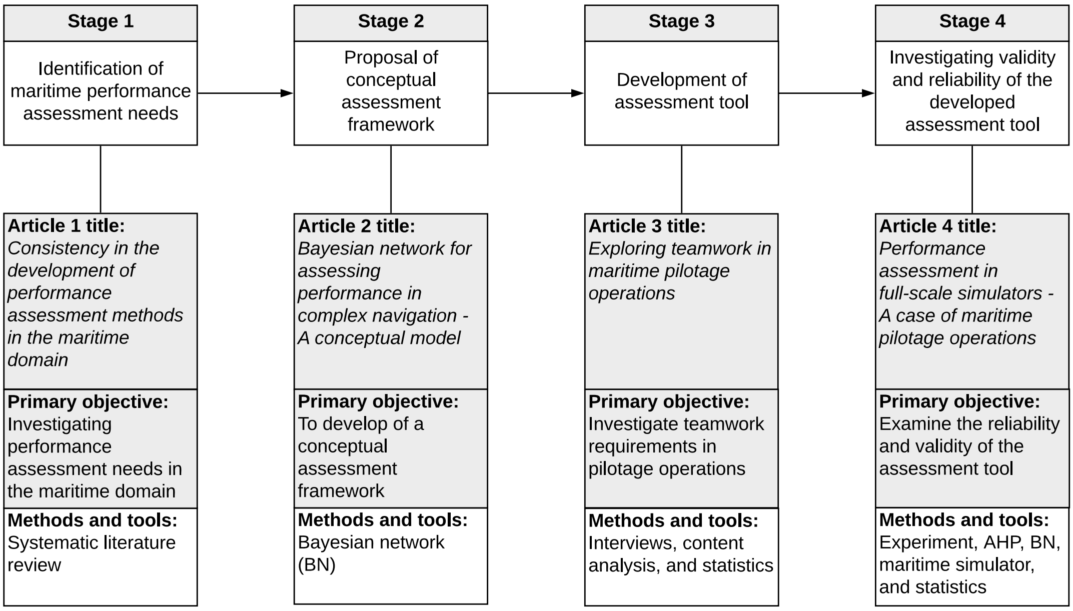
Cooperation		Shared mental models			
Team orientation	Mutual trust	Equipment	Job	Interaction	Team-knowledge
A propensity to take other's behaviour into account during group interaction and the belief in the importance of team goals over individual members' goals (Salas et al., 2005).	A shared perception that team members will perform actions relevant for all team members, and that the individual team members will recognise and protect the rights and interests of all the team members engaged in their joint endeavour (Simsarian Webber, 2002).	A shared understanding and knowledge of how to control technology and equipment (Converse et al., 1993).	A shared understanding of the task at hand, how to carry it out and environment's impact on their task (Converse et al., 1993).	A shared understanding of team member's roles and responsibilities within the team (Converse et al., 1993).	A shared understanding of team members knowledge, skills, abilities and preferences (Converse et al., 1993).

4 Methodological framework

This chapter provides an overview of the doctoral research program as well as details regarding each method that the researcher has used throughout the project. Section 4.1 presents the methodological overview of the project as a whole, then breaks down and explains the project's four stages. In Section 4.2, the central research methods are described. This section also presents an outline of the procedure for each of the appended articles and ends with a description of the tools comprising the CAPA-tool.

4.1 Overview of the doctoral research

To recap, this doctoral research project aimed to develop a tool that reduces the subjective impact in maritime performance assessment. This aim entailed a formalisation of the assessment framework. Different scientific methods were necessary to employ for achieving this aim; including literature reviews, collection of both qualitative- and quantitative data, in addition to various analysis tools of the respective data.



Stage 1	Stage 2	Stage 3	Stage 4
Identification of maritime performance assessment needs	Proposal of conceptual assessment framework	Development of assessment tool	Investigating validity and reliability of the developed assessment tool
Article 1 title: *Consistency in the development of performance assessment methods in the maritime domain*	**Article 2 title:** *Bayesian network for assessing performance in complex navigation - A conceptual model*	**Article 3 title:** *Exploring teamwork in maritime pilotage operations*	**Article 4 title:** *Performance assessment in full-scale simulators - A case of maritime pilotage operations*
Primary objective: Investigating performance assessment needs in the maritime domain	**Primary objective:** To develop of a conceptual assessment framework	**Primary objective:** Investigate teamwork requirements in pilotage operations	**Primary objective:** Examine the reliability and validity of the assessment tool
Methods and tools: Systematic literature review	**Methods and tools:** Bayesian network (BN)	**Methods and tools:** Interviews, content analysis, and statistics	**Methods and tools:** Experiment, AHP, BN, maritime simulator, and statistics

Figure 7: Overview of the PhD process.

The research progress is summarised in four stages, as illustrated in Figure 7 above. The boxes list the methods for each stage. Each stage has a different research focus, where the findings are forwarded over to the next stage. However, parallel work did occur even though the figure makes it appear streamlined. The scientific journey involved a need and opportunity of going back- and forth as more knowledge was acquired, although this figure does not capture this iterative process.

4.1.1 Stage 1 – identification of maritime performance assessment needs

The first stage of the doctoral research project was to conduct a literature review with the primary objective of identifying the performance assessment needs existing in the maritime domain. There were also two secondary objectives. One was to acquire an understanding of performance indicators in the maritime domain. Another secondary objective was to give the doctorate candidate a comprehensive overview of the research area.

Conducting a literature review has previously been identified as a useful contributor to this particular purpose for PhD students (Pickering & Byrne, 2014). The same work also identified that PhD students get an overview of theoretical- and methodological gaps in their respective research field, in addition to finding relevance and justification of their research. These benefits were central arguments for initiating the current doctoral work with a literature review.

The doctoral researcher published two articles for reporting the main objective in stage 1: The findings from the literature review were reported in the first appended article of this dissertation - Ernstsen and Nazir (2018a), in which a need for maritime navigation assessment was identified. Then, a conceptual framework for systematising performance indicators in the assessment of complex sociotechnical systems was developed and reported in an auxiliary article - Ernstsen, Nazir, Røed and Manca (2016). These findings and the generated knowledge carried over to the next research stage.

4.1.2 Stage 2 – Development and proposal of concept

Three different alternatives for performance assessment were considered before deciding which way to go. The first considered alternative was to develop a performance checklist using exploratory factor analysis. However, this alternative would require an infeasible large pool of specialised experts to tune the assessment tool properly. The second considered alternative, then, was to develop a fully automated assessment tool based on technical parameters, but this could be too rigid and cumbersome for practical use. Last, the preferred alternative was to develop an assessment tool based on theoretical and expert inputs organised in a structural probabilistic network: this was chosen due to its flexibility and forgiveness to a limited pool of experts.

The concept for a structural probabilistic network, which was a Bayesian Network (BN) for maritime navigation was developed in Stage 2 and reported in Article 2 – Ernstsen, Musharraf, Mallam, Nazir and Veitch (2018). The development of the concept required a comprehensive understanding of navigational pilotage needs (i.e., what to assess) and an understanding of the tool to be used for assessment (i.e., requirements for developing a BN).

Finding what to assess for pilotage operations required close collaborations with pilots and captains, e.g., Ernstsen, Nazir and Røed (2017) and Ernstsen & Nazir (2018b), as well as a comprehensive understanding of earlier literature on pilotage, e.g., Norros (2004) and Bruno & Lützhöft (2009). Moreover, to facilitate the learning of BN, the doctoral researcher was collaborating closely with a computer engineer specialised in the topic.

4.1.3 Stage 3 – Development of the assessment tool

As already mentioned, some parallel work occurred between the research stages. In stage 3, the focus was on the development of the tool itself. This focus required a formalisation of both the qualitative and quantitative features of the Bayesian assessment tool. The qualitative features comprise the theoretical understanding of the essential characteristics of a pilotage operation to assess. The main characteristics of pilotage were considered to be navigation, teamwork, berthing, getting the pilot onboard, as well as external factors such as traffic, weather, hydrodynamics and time pressure. Considering that navigation, berthing, getting the pilot onboard and external factors were investigated in Stage 2 of the project, Stage 3 aimed to explore the teamwork requirements for pilotage further. Although teamwork has been extensively covered in maritime human factors, i.e., in bridge resource management (BRM), studies on its effectiveness have failed to return results (O'Connor, 2011). It was then considered feasible to analyse teamwork requirements from the perspective of generic teamwork research (Rafferty et al., 2010; Salas et al., 2005).

Comprehensive interviews with pilots and captains were carried out for a top-down content analysis on teamwork in pilotage operations. The content analysis was based on a thorough review of teamwork in various complex operations, especially the work reported in Salas et al., (2005) and Rafferty et al., (2010) were significant inspirations for the content analysis. Moreover, the content analysis was an integral part of tailoring and adapting the teamwork literature for maritime pilotage operations. This work is presented in appended Article 3 – Ernstsen and Nazir (in second review).

The development of the quantitative features involved an investigation of how the various qualitative variables area weighted in relation to each-others. The weights were calculated using the Analytical Hierarchy Process (AHP) method, which is a multi-criteria decision-making method. Data on weighting the qualitative variables was collected in close collaboration with subject-matter experts.

Information on the qualitative and quantitative features comprise the assessment tool which was coined the Computer-Assisted Performance Assessment tool (CAPA-tool).

4.1.4 Stage 4 – Investigation of the CAPA-tool's reliability and validity

The final stage of the PhD focused on investigating the assessment tool's validity and reliability. More specifically, this stage involved carrying out experiments in which expert assessors would use the tool to assess maritime pilotage operations. Two approaches were considered during the design of the experiment. One was to run a high number of experiments and a low number of assessors or to design one experiment scenario that would be assessed by a higher number of assessors. While there are strengths and weaknesses with both approaches, the latter was deemed more feasible and practical.

The experiment scenario was designed, recorded, and edited before the running experiment. The design of the scenario made it possible to tailor it as needed. The standardisation of the assessment was also beneficial for the internal validity of the experiment. The scenario was presented in a remote assessment station prototype. The remote assessment station was a parallel research project. One main advantage of the remote assessment station was its portability, enabling the assessment of the maritime scenario outside of the full-mission bridge simulator environment.

The experiments were carried out soon after all preparations were completed. Participants were gathered from two Scandinavian maritime education institutions. The data collection required some travelling; however, it was highly valued to get inputs from at least two maritime educational institutions. Moreover, a small pool of participants was initially predicted, but the collaborative spirit from all participants was beyond expectations, and a (relatively) decent number of participants (n = 16) was acquired. When the experiments were completed, all responses were calculated for the respective experimental conditions. The results from the experiments, alongside a presentation of the assessment tool itself, is reported in appended Article 4 - Ernstsen and Nazir (in review).

4.2 Methods used in the doctoral research

A collection of various scientific tools was used in this doctoral research. The next sections describe the tools and necessary considerations for employing the tools in research.

4.2.1 Literature reviews

A literature review aims to give the reader a comprehensive understanding of the available literature (or knowledge) concerning a phenomenon. It is a popular method that continues to increase in number throughout various disciplines (Whittemore, Chao, Jang, Minges, & Park, 2014). There are different approaches to conducting a literature review, e.g., narrative reviews, meta-analyses, and systematic literature reviews; the latter was used in the current research.

The systematic literature review enabled the doctoral researcher to transparently and systematically define a research question, searching for relevant studies, assess its quality and then synthesise the findings (Armstrong, Hall, Doyle, & Waters, 2011).

4.2.2 Interviews

Interviews are an important and powerful tool for probing the opinions and perspectives of others (Myers & Newman, 2007). There are different methods to carry out an interview, in which the semi-structured interview technique is the most widely used in qualitative research (Willig, 2013). Semi-structured interviews give the researcher an opportunity to directing the conversation while the interviewee can speak freely about a particular aspect of the experience. At the same time, the researcher must be cautious to balance and maintaining control of the interview and where it is heading, while letting the interviewee have space to redefine the topic for generating novel insights for the researcher (Willig, 2013). This balancing is particularly important when eliciting information regarding expert decision making. A specific type of semi-structured interview for achieving this is the critical decision method (Klein, Calderwood, & Macgregor, 1989).

The critical decision method (CDM) provides a structure for eliciting expert knowledge (Klein et al., 1989). The method involves that the interviewer using specially designed probes to extract information pertaining to the interviewee’s cognitive processes during an incident, and, more specifically, the goals, strategies, and cues used (Morrison & Morrison, 2018). CDM can be summarised in five steps. 1) selecting an incident to be analysed, 2) gather and record information about the incident, 3) construct a timeline of the incident, 4) identify decision points and 5) ask questions around the decision points. One advantage of the technique is to enable experts to discuss similar incidents, although one must also acknowledge that the experts will have individual interpretations of the incident.

The critical decision method technique was adapted to elicit information about important characteristics of pilotage operations in Stage 2 of the PhD. Moreover, to improve the validity of the interviews, the interviewer strived to develop a good rapport with the interviewees. This was done by following the PEACE technique, as suggested in Clarke and Milne (2001, p. 187).

The PEACE technique is a mnemonic for planning and preparing, engage and explain, account, closure, and evaluation (of the interview and the interviewer's performance). The "P" entails that the researcher can demonstrate a good knowledge of the topic and has planned how to conduct the interview, e.g., having planned a proper icebreaker to start building rapport. The "E" suggests that the interviewer must introduce him- or herself properly, explain the purpose, and ensures that the interviewee gives informed consent; that is, ensure that the interviewee knows his- or her rights concerning data protection and anonymity. The "A" is about carrying out the interview questions correctly, maintain the interview to relevant topics, and the use of pauses, silence, and body language as part of active listening and communication. "C" is regarding finalisation of the interview, summarises the conversation and invites the interviewee for final comments, as well as allowing him- or her to add, alter and correct information. The final "E" is that the researcher evaluates strong and weak points with carrying out the interview. The PEACE technique served helpful for the doctoral researcher in ensuring quality when conducting interviews.

4.2.3 Experiments

Experimentation is a method for collecting empirical evidence for cause-and-effect relationships by manipulating certain factors (Campbell, 1963). An experiment must carefully balance the amount of control and the possibility of generalising the results out of the laboratory. Often, higher internal validity is achieved on account of external validity (especially ecological), and vice versa (Brunswik, 1956; Messick, 1987).

This balance was essential to the experiment carried out in stage 4 of the doctoral project considering the difficulty of having control of all factors in the experiment. The purpose of the experiment was to examine the reliability and validity of the assessment tool and is reported in Article 4.

4.2.4 Data analysis

4.2.4.1 Content analysis and coding procedure

Content analysis has become increasingly popular and has a long history of use, particularly in the fields of psychology, sociology, business, journalism and communication (Neuendorf, 2016, p. 27). There are several variations of content analyses; however, a typical content analysis procedure can be: theory and rational (i.e., what will be examined and why?), conceptualisation (i.e., which variables are you looking for?), operationalisation (i.e., how will you measure the variables?), coding schemes (i.e., how content will be coding), sampling (how data is collected), coding (coding the data according to the coding scheme) and finally, an estimation of interrater reliability (calculating the reliability figure). Neuendorf (2016, pp. 50–51) provides a comprehensive process for this method. Furthermore, a method of coding the data is to break transcribed statements down into meaningful units and condensed meaningful units; then the meaningful units are defined as a category (Graneheim & Lundman, 2004).

4.2.4.2 Interrater reliability

Two methods different are employed in the current study for investigating this: the Cohen's kappa (Cohen, 1960) and the Krippendorff's alpha (Krippendorff, 2011b, 2011a).

Cohen's kappa and Krippendorff's alpha are two varieties of calculating the agreement between raters while controlling for the impact of chance agreement. Cohen's kappa is reported as the most widely used reliability coefficient, whereas the Krippendorff's alpha statistic is considered highly attractive, but less used due to its tedious calculations

(Neuendorf, 2016, pp. 150–151). Two different methods was used since the kappa statistic assumes nominal-level data and derive from the relationship between the proportion of observed agreement and proportion of agreement that can be expected from chance (Cohen, 1960). Krippendorff's alpha, however, also takes into account the magnitude of the misses by looking at the observed disagreement and expected disagreement and can be used with all measurement scales (Krippendorff, 2011a).

4.2.5 Tools used in the CAPA-tool

This section presents the analytical hierarchical process (AHP) and the BN. The AHP was used to weigh the factor's count on each of the performance indicators, whereas the BN was used to structure and calculate the performance score for all levels of the hierarchy (based on the weights deriving from the AHP). Both methods have been widely used for this purpose earlier, e.g., Podgórski (2015) that demonstrated an AHP-based selection of performance indicators and Musharraf, Smith, Khan, Veitch and MacKinnon (2016) that used a Bayesian example of assessing evacuation behaviour. Additionally, Millán, Descalço, Castillo, Oliveira and Diogo (2013) details how BNs can be used to improve knowledge assessment. The next sub-sections expand further on these two methods as they are central to the CAPA-tool.

4.2.5.1 Analytical hierarchical process

AHP is a tool in which paired comparisons derive ratio scales among different choices and criteria. This comparison method can be used to calculate a weight for respective assessment criteria across a multitude of experts; thus, represent an aggregated understanding of how criteria should be weighted. AHP was initially developed by as a structured technique for analysing complex decisions (T. L. Saaty, 1988). It can be used to rank, prioritise, weigh, allocate resources, provide benchmarks, quality management, and conflict resolutions. It is difficult to represent and generate quantitative indicators based on qualitative data, but AHP is considered a viable method for measuring and quantifying subjective opinions without too much of a compromise (Manca et al., 2014; R. W. Saaty, 1987).

The AHP procedure can be broken down into segments which involves modelling the problem as a hierarchy with a decision goal, then establish priorities among the elements based on judgments, checking the consistency of judgments; finally taking a decision based on the numerical result (T. L. Saaty, 1990). The hierarchy should attempt to capture the complexity of the decision, yet nimble enough to be sensitive to changes. Please see Figure 8 as an overview and the three steps of AHP below.

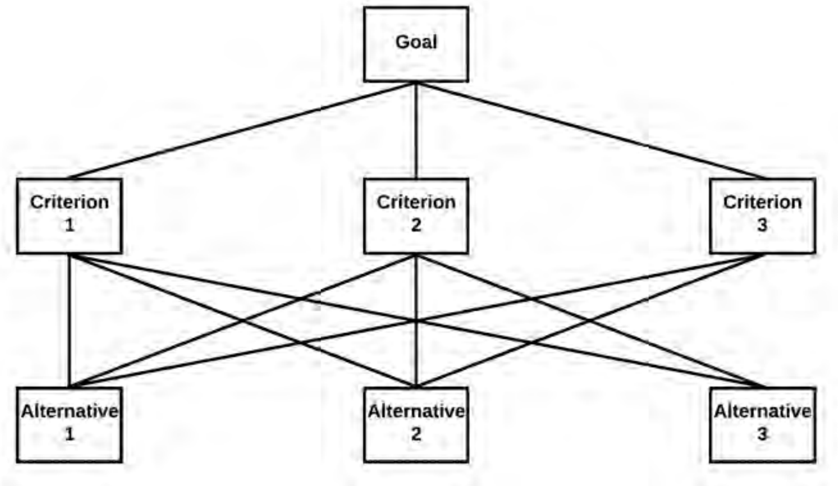


Figure 8: Criterion modelling in AHP.

1. Computing the vector of criteria weights: The AHP starts by creating a pairwise comparison matrix. T. L. Saaty (1977) suggested a fundamental scale with five classifications set to values of 1 to 9 (1, 3, 5, 7, and 9) to be used during the evaluation in constructing the matrix as shown below in Table 8 below.
2. Computing the matrix of option scores: The pairwise option evaluations are performed by comparing the values of the performance indicators corresponding to the decision criteria. Hence, this step of the AHP can be considered as a transformation of the indicator matrix into the score matrix.
3. Ranking the options: Once the weight vector and the score matrix have been computed, the AHP obtains a vector of global scores by multiplying the *weight vector* and *score matrix*. As the final step, the option ranking is accomplished by ordering the global scores in decreasing order.

Table 8: List of weights criteria.

Value	Definition	Explanation
1	Equally importance	Identical contribution
3	Weak importance	Slightly superior judgment
5	Strong importance	Strongly judgment in favour
7	Very strong importance	Recognized dominance
9	Absolute importance	Confirmed dominance
2,4,6,8	Intermediate values	When compromise is needed

4.2.5.2 Bayesian network

BN is a probabilistic model that represents an interaction of random variables through a directed acyclic graph and conditional probability tables (Pearl, 2014). The relationships between the performance indicators are represented as a combination of nodes, and the relationship between the nodes are represented as arcs. A BN consists of a qualitative and quantitative component. The graphical representation of the network is the qualitative composition.

The qualitative composition of a BN can be developed through theoretical input, task analyses or other forms of meaningful input. It graphically represents the operation where one node can be attributed to a significant element in the operation, for instance a main task. Its parent and child task can thus represent the task's function or other subtasks. The qualitative aspects of the Bayesian network are subjective and is prone to limitations associated with subjective research methods, e.g., observation bias, expectancy bias and selection bias. The graphical representation, however, also requires a quantitative probability distribution.

The quantitative distribution focuses on the variables' associated probabilities. In the network, two types of probabilities must be quantified: prior probabilities of the independent variables (i.e., root nodes, empirical indicators, initial criteria) and the conditional probabilities of the dependent variables. This probability specifies the probability of each child nodes (dependent variable) for every state of its parent (directly dependable variable).

Information about these variables allows us to calculate the probabilities of the network's child nodes. If there are *n* variables $X_1, X_2, \ldots, X_n$ in the network and *Pa*(X_i) is the set of parents of each X_i, we can calculate the joint probability for the entire network using eq. 1, in which the discrete conditional probability of X_i given its parents is *P*(X_i |*Pa*(X_i)).

$$P(X_1, X_2 \ldots\ , X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)) \qquad (1)$$

The quantitative part can be subjective and objective (e.g., based on statistics or weighted using multi-criteria decision-making tools, such as AHP). This combination of a qualitative construction and a quantitative and objective feature of the method makes it a versatile tool for assessment purposes. Please see Figure 9a and 9b below for a simple conditional probability table (a) and structure (b).

	Yes	No
EI1	1	0
EI2	0	1

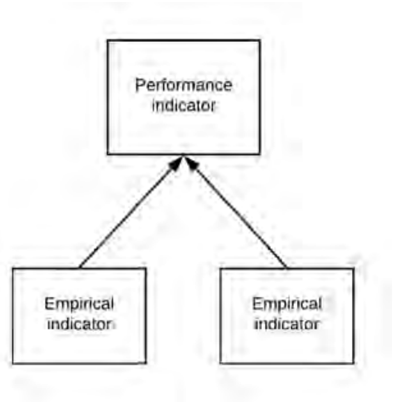


Figure 9a and 9b: a) A simple conditional probability table for a performance indicator. In this instance, the empirical indicator 1 is active and 2 is not. b) A simple representation of a Bayesian assessment hierarchy. Empirical indicators are observable variables, whereas the performance indicators are higher-order non-observable indicators. More levels to the hierarchy can be found in complex representations.

Bayesian models are commonly used to model various assessment processes, e.g., Millán et al., (2013) and Musharraf, Hassan, Khan, Veitch, MacKinnon and Imtiaz (2013). There are many advantages of using the method of modelling. It is suitable for small and incomplete data sets. It is also viable for continuous development by improving its reliability, resolution and uncertainty (i.e., its brier score) by continuously updating the probabilities as more data becomes available over time.

4.3 Procedures

4.3.1 Article 1 - procedure

4.3.1.1 Research statement, database search, and exclusion criteria

The research statement for Article 1 was “the use and development of performance indicators in the maritime industry”. The statement was broken down into four concepts to be used as keywords when searching the literature. The four concepts were “Performance indicators”, “Maritime”, “Framework”, and “Method”. Different keywords relating to the concepts were identified and added to the Boolean search string:

(“Performance indicators” OR “key performance indicators”) AND (maritime OR marine) AND (framework OR measure OR reference model) AND (method OR methodology) AND shipping).

The string returned 537 (after removing 193 duplicates) distinctive papers from Scopus, ScienceDirect, and JSTOR. A formalised exclusion process was developed for systematising the review. The first exclusion criterium was to investigate whether the respective papers were addressing the maritime domain. It was necessary as the majority of papers triggered by the key-string concerned ecological- and marine biological studies; not directly tied to the shipping industry. This process was carried out by investigating the papers’ abstracts. Following the first criterium, 128 papers were selected for further study.

Next, the study also disregarded conceptual papers where the arguments were inadequately supported theoretically- or methodologically. This process required that the researcher had to sift through all of the papers. In parallel, criterium 3 was also considered: whether the performance assessment covered in the respective papers associated at the operational or tactical level of analysis. This criterium was necessary for excluding papers interested in organisational performance assessment, such as the

ShippingKPI index (Sleire & Dale, 2009). Sixty-two papers remained after the entire exclusion process.

4.3.1.2 Analysis

The remaining papers were then cross-referenced in two dimensions: 1) their approach used for the development of the assessment (e.g., a top-down approach) and 2) which shipping domain the paper was related to (e.g., ship-handling). The analysis consisted of univariate- and bivariate analyses. The univariate analysis was performed to determine the distribution of the data. The bivariate analysis was a cross-tabulation of the two dimensions. Then, the assessment approaches were scored (based on expert reviews), and their relative score (for normalising the data) were calculated and compared among the four domains.

4.3.2 Article 2 - procedure

Article 2 was a concept paper and thus, no methodological procedure to present here. However, it is worth mentioning that a central element in the development of the paper was learning how a BN for assessing complex operational performance could be used. This process entailed applying a BN to generic complex navigation. Also, an auxiliary paper to this dissertation used a similar approach to investigate challenges on the bridge during pilotage navigation (Ernstsen, Musharraf, & Nazir, 2018).

This learning was an important foundation for the development of the conceptual assessment tool. The development required close collaboration with computer engineers possessing an expertise in BN. The joint papers summarise the ideas generated during this work.

4.3.3 Article 3 - procedure

Ten interviews were conducted. The participants were recruited to represent both pilots and captains equally. They were recruited through networking and the “snowball method”, in which the participants themselves could suggest colleagues for subsequent interview recruitment. The mean age of the participants was 47 years (SD = 8.9), and their mean years of sailing was 19.3 years (SD = 14.4). Their sailing experience was dispersed among oil ships, cargo ships, and cruise ships. All ships had sailing connections to Scandinavian ports, and the participant had frequent encounters with sailing in these waters.

The interviews were carried out using the critical decision-making approach, adapted from Klein et al., (1989). Moreover, all interviews were conducted by following the PEACE technique (i.e., Preparing, Explaining, Accounting, Closing and Evaluating the interviews), as suggested in Clarke and Milne (2001, p. 187).

The interviews were semi-structured with a focus of being conversational and non-confrontational, in addition to using probes for gathering data. This method had two benefits in this study. One, it enabled the interviewer to ensure that the required data was collected by expanding on the central topics that the interviewee put on the table. Two, the interview method allowed the researcher to gather data exploratively, compared to other more structured approaches to interviews (Willig, 2013).

All interviews lasted 1 hour and 4 minutes on average. The most extended interview was 1 hour and 34 minutes and the shortest only 52 minutes (see Table 9 below for individual interview data). The data was audio recorded and securely stored with permission from the National Centre for Research Data (project number: 51322) and with interviewee's informed consent. Furthermore, the interviews were mostly collected in-person, two were conducted using FaceTime® (an alternative video telephone service to Skype™). While in-person interviews usually give more abundant data, video-telephone interviews are considered a viable compromise when time- and geographical constraints make in-person interviews impractical (Rowley, 2012). Finally, all data were transcribed – first verbatim, then pragmatic as the data collection was becoming saturated – and then organised, translated and prepared for subsequent analyses.

Table 9: Participant role and interview length.

ID	Role	Interview length
1	Pilot	1 hour and 34 minutes
2	Captain	52 minutes
3	Pilot	57 minutes
4	Captain	1 hour and 7 minutes
5	Pilot	1 hour and 13 minutes
6	Captain	53 minutes
7	Pilot	1 hour and 11 minutes
8	Captain	54 minutes
9	Pilot	1 hour and 6 minutes
10	Captain	56 minutes

4.3.3.1 Data analysis

The data analysis consisted of a qualitative and a subsequent quantitative part based on the finding from the analysis of the qualitative interview data. The qualitative approach was a content analysis and the quantitative consisted of inferential statistics.

The interview data were analysed using a deductive content analysis approach. The deductive element entails that the interview data is analysed following a theoretical framework, as opposed to the bottom-up approach, which grounds the analysis in the data. The content analysis consisted of the steps following the renown Neuendorf (2016). The steps that were following in this study are 1) research question, 2) conceptualisation, 3) operationalisation, 4) coding procedures, 5) data sampling, 6) coding, and 7) interrater reliability. Please see Table 10 for an overview of the procedure and measures that were taken for each of the steps.

Table 10: The content analysis procedure and actions taken for each procedure.

Step	Procedure	Action
1	Aim of study	To examine the core teamwork factors in the case of a maritime pilotage operations.
2	Conceptualisation of teamwork	The core factors representing teamwork in the current study are based on the conceptualisations deriving from the literature reviewed in Rafferty, Stanton & Walker (2010).
3	Operationalisation	Four variables are operationalised as eleven subfactors (see Section 3.2).
4	Coding scheme	The coding procedure followed Graneheim & Lundman's (2004) process in which each statement was broken down into a condensed meaning unit, then a condensed interpretation before assigning a code that best fitted that particular statement.
5	Data sampling	A census of the content was possible: all statements transcribed from the interviews are included in the analysis.
6	Coding and interrater reliability	The interrater reliability process followed the guidelines proposed in Kottner, et al., (2011) and Lombard, Snyder-Duch, & Bracken (2002). One coder conducted 100 % (n = 210) of the analysis and two independent reliability coders shared the coding of 24 % (n=50) of the statements to assess the interrater reliability.
7	Final reliability	The interrater reliability agreement was calculated using Cohen's Kappa (Cohen, 1960) in IBM SPSS version 25.0.

A critical step in the content analysis is the coding procedure; that is, the way information is generated from the interview data. The coding procedure in the current content analysis followed Graneheim and Lundman (2004), in which each interview statement was broken down into condensed meaning units, condensed interpretations, and then into a code that were considered to fit accordingly. Please see Table 11 below for an example of the coding procedure:

Table 11: Example of coding procedure.

Statement	Condensed meaning unit	Condensed interpretation	Category	Sub-category
"Ideally, I ask questions to the captain regarding data on the ship, like what type of rudder etc. "	Pilot ask bridge about ship specs	Pilot ask for information about ship specifications	Communication	Information exchange

Finally, the interrater reliability (agreement) was calculated using Cohen's Kappa (κ). It was derived by using eq. 2, where PA_O is the observed per cent agreement and PA_E is expected (by chance) per cent agreement (Cohen, 1960). Cohen's kappa is reported as the most widely used reliability statistic (Neuendorf, 2016, p. 150). It ranges from 0.0 (agreement at chance level) to 1.0 (perfect agreement).

$$\frac{PA_O - PA_E}{1 - PA_E} \tag{2}$$

4.3.4 Article 4 - procedure

4.3.4.1 Data collection

Expert navigators (n = 16) from different maritime academy institutions in Scandinavia were recruited for the experiment. Sailing, simulator and assessment experience were the criteria for participating in the experiment. These criteria were necessary for ensuring that the participants possessed the prerequisite expertise. All participants were placed pseudo-randomised in either the experiment condition or the control condition. This placement procedure was to make sure that both conditions were equally represented. However, it was random which particular participant that was placed in the separate conditions. The pseudo-randomisation was necessary due to the specialised and limited pool of participants. All participants received and signed an informed consent form explaining the purpose of the study and their rights as participants concerning data protection and handling. The National Centre for Research Data (project number: 181630) approved the collection and storage of relevant research data.

There were two experimental conditions: the control group and the experiment group. The control group was using a conventional questionnaire for assessing the navigation scenario, adapted from a professional assessment course. The participants in this group were asked to derive a performance score from 0-100 for the pilot-bridge team's teamwork performance, technical performance and one for their overall performance. The experiment group was using the formalised assessment tool developed in this current doctorate research. The participants in this condition were not asked to derive a performance score as this is calculated directly by the CAPA-tool based on the assessor's input.

The K-SIM™ simulator was used to record a maritime pilotage scenario and to prepare the experiment. It is a full-scale maritime navigation simulator consisting of six widescreen monitors that imitate the view of the bridge, including realistic instruments and panels (the hardware) in a 1:1 aspect ratio. This full-mission bridge is designed to preserve a realistic and immersive representation of navigating on a ship's bridge. All necessary tools were available for the participants, such as the rudder-, throttle-, thruster controls.

Then, the recorded exercise was exported and integrated into a remote assessment station. The assessment station was designed to enable assessment of the navigation scenario external to the proprietary simulator software. This modification was necessary to ensure portability and standardisation of the assessment environment, thus making it convenient for multiple raters to assess the scenario. The assessment station was created in close collaboration with expert simulator assessors (who subsequently was not part of the experiment in consideration of internal validity, such as systematic errors). Please see Figure 10 below for the setup of the video screens.

The experiment began with briefing the participant (the same procedure for both of the groups). They were allowed to ask questions for further clarifications of expectations concerning their assessment tool and the overall purpose of the assessment. They were shortly after given the informed-consent form with essential elements about their data and anonymity considerations.

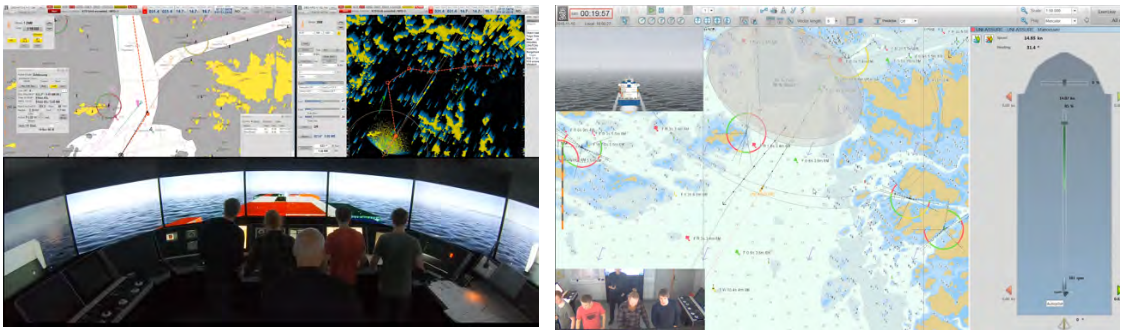


Figure 10: Setup of the video screens (Integrated from Article 4).

The participant began the exercise using their respective assessment tools shortly after the briefing. The assessment scenario lasted for 1 hour, 27 minutes and 25 seconds. Due to the length, the participants were given freedom regarding coffee- and bathroom breaks. Upon completion, all participants received a debrief where they were allowed to discuss their experiences of assessing the scenario. At this point, all relevant data for this experiment was collected so it was considered to be unproblematic for the internal validity to conduct this debrief. Please see Figure 11 below for an outline of the experiment process.

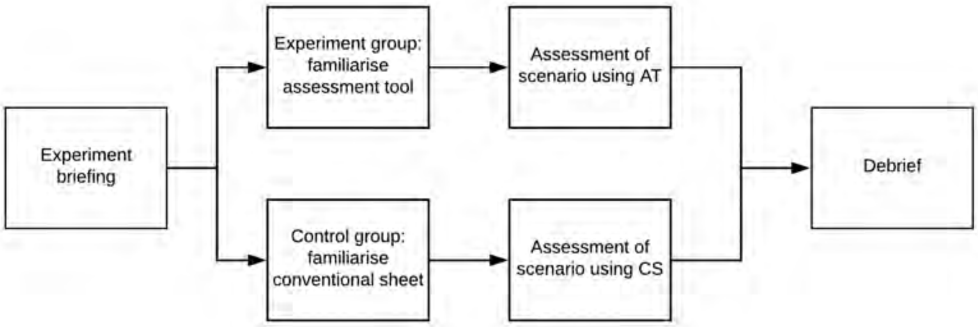

Experiment briefing
Experiment group: familiarise assessment tool
Control group: familiarise conventional sheet
Assessment of scenario using AT
Assessment of scenario using CS
Debrief

Figure 11: Outline of experiment process (integrated from Article 4).

4.3.4.2 Data analysis

Three different statistical analyses were conducted for information concerning the reliability and validity of the assessment tool: Krippendorff's nominal alpha, coefficient of variation (CV), and a one-sample t-test.

Krippendorff's nominal alpha computes the interrater reliability among the raters' responses to the assessment tool. It was calculated using an online calculation tool – the ReCal tool (Freelon, 2011). The alpha was calculated using eq. 3, where D_O is the observed disagreement, and D_E is the expected disagreement.

$$1 - \frac{D_O}{D_E} \tag{3}$$

The tool's absolute reliability was investigated by comparing the variation between the experimental conditions for the three dimensions: technical score, teamwork score and total score. Absolute reliability pertains the degree to which repeated measurements vary, whereas relative reliability pertains the degree an individual maintains his or her position over repeated measurements (Safrit & Wood, 1989, pp. 45–72). The variation was standardised as a coefficient and can be interpreted as the dispersion among raters;

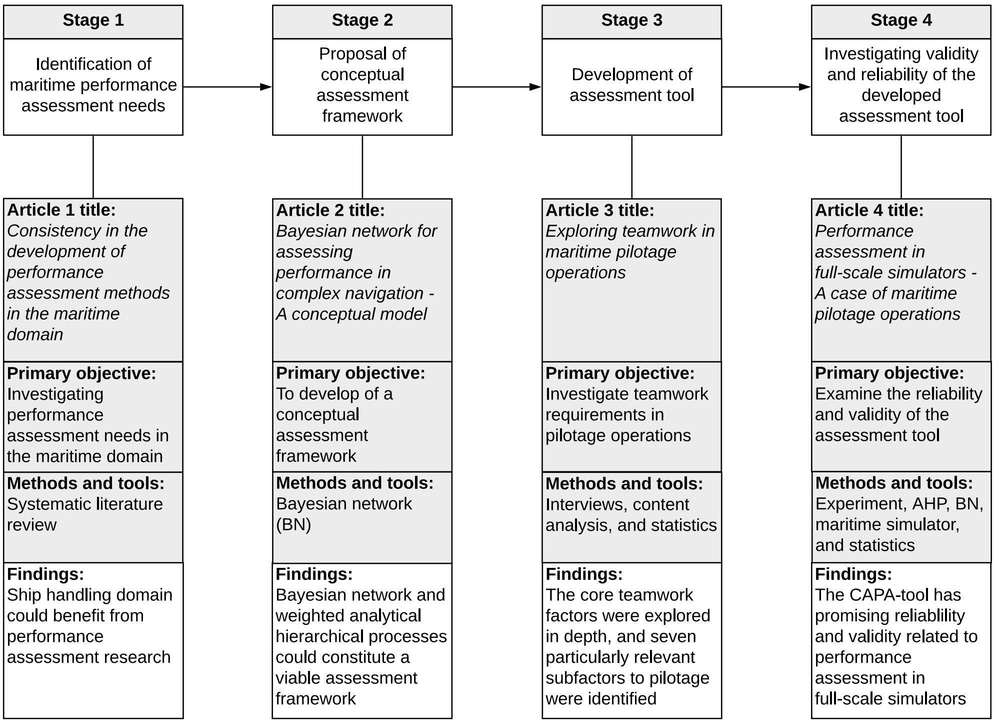
a lower percentage score is associated with less dispersion between individuals. It should be clarified that the conventionally employed intraclass-reliability coefficient (ICC) was improper for this particular dataset as it is a function of both the within- and between rater variations (Atkinson & Nevill, 1998).

The work in Article 4, however, was also centred around examining and collecting evidence of construct validity. Ideally, this evidence is accumulated from numerous studies on the measuring instrument. In Article 4, this was investigated by correlating the performance scores to a gold standard that was designed to represent a “true” performance of the trainees. Although, ideally, a comprehensive study to collect data for a factor analysis would have been conducted, though this would be infeasible considering the specialised sample required to run this study.

One-sample T-Test. This is a statistical procedure for determining if a sample of observations could have been generated by a process of a specific mean. It was carried out to investigate if the average assessment score could statistically be connected to a golden assessment standard. The test, thus, analysed if the data was statistically different from the gold standard. Moreover, the small sample size and failure to meet normality suggests careful interpretations of the *t*-value; therefore, it comprise an elevated risk of committing a type 2-error (De Winter, 2013).

5 Results

The results reported in the four appended articles are presented in this chapter. The main findings in relation to the four stages are summarised in Figure 12 below. Article 1 reports a literature review in which assessments needs in the maritime domain are identified. This finding initiated the development of a conceptual assessment framework, which was reported in Article 2. The learning from the conceptual stage is forwarded to the development of the tool itself. Further research that was required, was to understand the teamwork needs in pilotage operations. Article 3 reports this scientific examination, where the sub-factors of communication, coordination, cooperation, and shared mental models were found to be invaluable aspects of pilotage operations. The information collected in stages 1-3 was then used to develop a computer-aided performance assessment tool (the CAPA-tool) for training in full-scale simulators. Article 4 reports the experiment where it was found good absolute reliability in the assessment of technical competencies, although the reliability for assessing teamwork was lacking. The important findings are detailed in the remaining parts of this chapter.



Stage 1	Stage 2	Stage 3	Stage 4
Identification of maritime performance assessment needs	Proposal of conceptual assessment framework	Development of assessment tool	Investigating validity and reliability of the developed assessment tool
Article 1 title: *Consistency in the development of performance assessment methods in the maritime domain*	**Article 2 title:** *Bayesian network for assessing performance in complex navigation - A conceptual model*	**Article 3 title:** *Exploring teamwork in maritime pilotage operations*	**Article 4 title:** *Performance assessment in full-scale simulators - A case of maritime pilotage operations*
Primary objective: Investigating performance assessment needs in the maritime domain	**Primary objective:** To develop of a conceptual assessment framework	**Primary objective:** Investigate teamwork requirements in pilotage operations	**Primary objective:** Examine the reliability and validity of the assessment tool
Methods and tools: Systematic literature review	**Methods and tools:** Bayesian network (BN)	**Methods and tools:** Interviews, content analysis, and statistics	**Methods and tools:** Experiment, AHP, BN, maritime simulator, and statistics
Findings: Ship handling domain could benefit from performance assessment research	**Findings:** Bayesian network and weighted analytical hierarchical processes could constitute a viable assessment framework	**Findings:** The core teamwork factors were explored in depth, and seven particularly relevant subfactors to pilotage were identified	**Findings:** The CAPA-tool has promising reliablility and validity related to performance assessment in full-scale simulators

Figure 12: Overview of the PhD process with additions of the findings.

5.1 Article 1 – findings

The literature review investigated the maritime domain's use of performance assessment methods. It was found using a bivariate analysis that ship handling (defined in the article as manoeuvring and navigating the ship) returned the lowest relative score concerning the development of assessment methods. The highest scores were connected to research on safety and environment, whereas logistics research returned a mediocre score relative to the three other domains. Please see Table 12 below for a summary of the findings.

Table 12: Bivariate analysis of assessment approach and maritime domain.

Approach	Weight (w)	Port logistics	Ship handling	Safety	Environmental
Bottom-up	1.0	4	3	0	2
Top-down	1.0	8	4	2	7
Hybrid	1.5	10.5	0	6	3
Inadequate	-0.5	-5	-3	-1	-0.5
Score		17.5	4	7	11.5
Maximum score		43.5	19.5	12	18
Relative score		0.40	0.21	0.58	0.64

In the table above, the two dimensions are cross-referenced: maritime domain and the approaches used for developing assessment methods. The number of papers using the various approaches (i.e., bottom-up, top-down, hybrid, inadequate) were associated with the four domains (i.e., port logistics, ship handling, safety, environment). Moreover, a relative score was calculated to correct for the number of papers returned in the literature review for the respective domains. This normalisation enabled the researchers to compare the results more easily. While it was meaningless to draw conclusions based on the maximum score, the relative score can at least suggest which of these four domains that call for attention concerning the development of assessment tools.

The finding that the ship handling domain could benefit from performance assessment initiated the development of a performance assessment tool. This inspiration was a key component when moving the project into Stage 2 – development of a conceptual assessment tool.

5.2 Article 2 – findings

The concept published in Article 2 was central in the development of the assessment tool. It was a result of strenuous research concerning theoretical and technological demands associated with the assessment of maritime navigation. While several approaches were considered, certain design requirements were identified in light of the theoretical framework and delimitations presented in chapters 2 and 3. Some of the considerations will be provided here.

As previously discussed, two critical elements for the assessment tool were to reduce the subjective impact of assessment and to maintain flexibility. Reducing the subjective impact implied a need for pre-articulated and pre-weighted criteria based on inter-subjective enquires. This formulation of criteria enables assessors to make judgments without having to self-determine how the various elements should be weighted. This judgment call is further transferred to a computerised calculation that is not (directly) prone to human bias.

However, achieving this while maintaining assessment flexibility insinuated that the human assessor must be involved. Thus, the human operator is maintained as a sensor, assessing the performance without determining each factor's weight on the overall performance score. This arrangement makes it possible to use the assessment tool across different types of operations. An assessment tool based on the BN and weighted using AHP could therefore maintain flexibility while enabling the human assessor the opportunity to provide observational input, like a human sensor. Another key element was the potential of giving diagnostic feedback to the trainees. The standardised assessment tool returns data concerning all variables in the different scenarios. This feedback enables the trainee to evaluate his or her performance and to make subsequent adjustments to further improve his or her competency.

The concept was integral to development of the CAPA-tool.

5.3 Article 3 – findings

Earlier investigations revealed that numerous publications were already related to tasks involved in pilotage operations; thus, Stage 3 concentrated on exploring teamwork in piloting operations. Four core teamwork factors and eleven subfactors were examined in this stage. The results from the interrater reliability study will also be presented, as well as frequency statistics on the number of statements per subfactor.

5.3.1 Interrater reliability and frequency statistics

The proportion of joint judgments in which there is agreement was moderate (k = .706) after chance was excluded (McHugh, 2012). The interrater reliability was, however, considered acceptable due to the operational complexity described above. For the analysis, 50 statements (25.4 % of n = 197 statements in total) were randomly sampled and divided between two raters (60/40 % split).

This analysis following the size suggestion of Lombard, Snyder-Duch and Bracken (2002) stating that the reliability sample should either contain a minimum 50 statements or 10 % of the full sample. The reliability coders were independent researchers with relevant expertise, and they were also trained for 30 minutes before starting the reliability assessment. A second expert (a co-author not involved in the coding procedure) functioned as the tiebreaker for the 13 statements where the interraters disagreed.

Furthermore, all core teamwork factors (communication, coordination, shared mental models and cooperation) were recognised in the interview material. However, the content analysis also revealed anomalies about the subfactors when applied to the case of maritime pilotage. This deviation from the standard literature was particular for adaptability, in addition to three of the four subfactors of shared mental models.

These findings can also be echoed in the bar graph below that depicts the number of statements coded to each of the subfactors (Figure 13). In this graph, adaptability, equipment, interaction and team-knowledge are all less covered than the other subfactors.

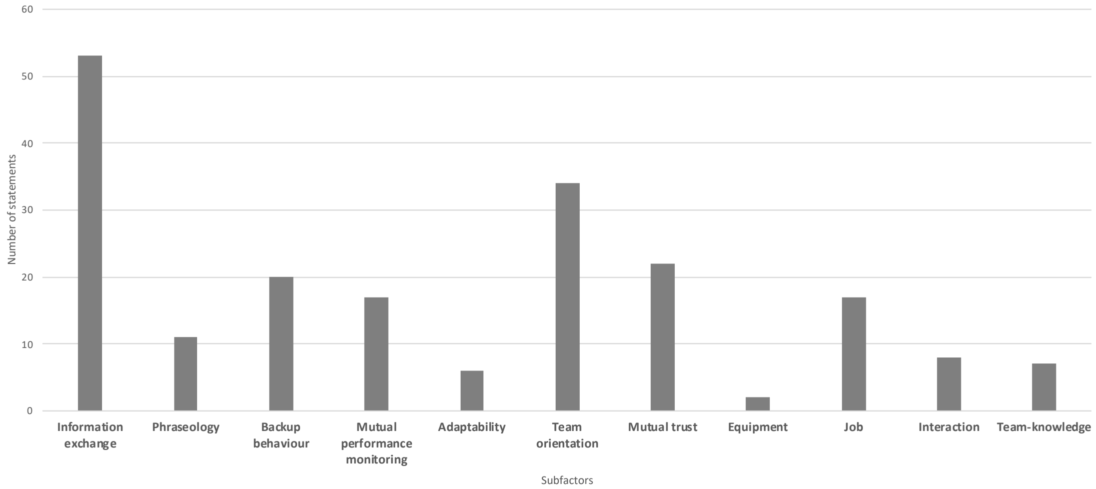


Figure 13: Bar graph showing the number of statements that were coded for each of the subfactors.

5.3.2 Communication – key findings

Effective information sharing was found to be an indispensable part of pilotage. Particularly critical moments of information sharing were related to the initial stages of the operation, in which the pilot first enters the bridge. The pilot and the captain must in this early phase exchange information about the ship's technicalities; such as its draft, any malfunctions and similar information. They must also share information about the surroundings; for example, information about any peculiar traffic conditions or hidden skerries that the bridge team must know.

It was also pointed out that an active exchange of information during navigation is also vital to the operation. This activity includes giving commands for speed and course, information about the port (e.g., berth positions, which cranes the ship needs to connect to). It also includes information exchange with external agents, such as vessel traffic services, surrounding ships, tugboats, mooring crew and other pilots (often referred to as colleagues, but also friends).

When discussing how information should be shared – the phraseology – it was repeated that it was more important that information was shared, than how it was shared. In fact, some participants pointed out that they were using their private cell phones to connect to external agents, although mostly with other pilot colleagues, for more information on the operation. A similar philosophy of deviating from protocol was mentioned in relation to the use of pilot cards. It was emphasised that the circumstances would dictate whether the pilot should review the pilot cards. If the need for navigational aid was immediate, the pilot would ask to instead receive the pilot cards at a later point in time, when the situation would be more predictable. Please see Table 13 below for example statements.

Furthermore, it was mentioned that this decision could result in a dichotomy, and even a goal conflict, between the captain and the pilot, i.e., one of the agents wanted to focus on the operation and the other wanted to follow protocol and assess the pilot cards. In light of this discussion, it was apparent that the attitudinal component (cooperation) of teamwork in the pilot-bridge team was imperative for effective exchange of information regardless of the specific phraseology.

Table 13: Example statements coded as communication.

Interview	Statement	Subfactor
Pilot	*"When the pilot is boarded, he will notify the VTS (i.e., vessel traffic services) about his arrival and his route intentions and also asks for information about surrounding traffic."*	Information exchange
Captain	*"When the pilot enters the bridge, there will be an information exchange between me and the pilot. He usually asks for maximum drafts and stuff like that."*	Information exchange
Pilot	*"When the situation is under control after boarding, I get vessel information communicated both orally and through pilot cards."*	Phraseology
Captain	*"We use closed-loop communication with the pilot during navigation to make sure that the commands are understood."*	Phraseology

5.3.3 Cooperation – key findings

Team orientation and mutual trust were extensively discussed, 34 and 22 statements respectively. Both the captains and pilots emphasised the need to facilitate team orientation in the early stages of the pilotage operation. The interviewees discussed elements such as conversational ice breakers, serving coffee or lunch to the pilot, as well as greeting the pilot properly, as crucial for the development of team orientation. These elements were considered important for building *rapport* between the key agents in the bridge team.

Furthermore, strategising and planning were also considered to be essential characteristics of team orientation. The pilot and the captain must plan the operation early, which encompasses that they determine and decide which route to take and distribute responsibilities among key members of the team (especially for critical phases of the operation). The distribution of responsibility could, for instance, be to decide who will be steering the ship during berth. The interviewees articulated that it was essential that these issues were clarified.

In similar terms, mutual trust was considered important. The participants stated that the pilot-bridge team must strive to have a positive team chemistry on the bridge. It involves that the team-members trust that the other team-members do what they consider is the best for the team. A critical feature of having proper mutual trust on the bridge was the issue of information sharing; that is, proper trust inclined the members to also share nice-to-know information, as opposed to strictly sharing need-to-know information. Please see Table 14 below for example statements.

Table 14: Example statements coded as cooperation.

Interview	Statement	Subfactor
Pilot	*“If there is a positive bridge culture, I will strive to ensure that the helmsman or captain have understood my advice.”*	Team orientation
Captain	*“I (the captain) like to plan with the pilot who's doing what, who's sailing this and this leg and who's berthing.”*	Team orientation
Pilot	*“An important role for the pilot is to generate trust and good chemistry in the group.”*	Mutual trust
Captain	*“It's very important that the crew and I all accepts and welcomes operational feedback, it’s very easy to miss small, but important navigational considerations.”*	Mutual trust

5.3.4 Coordination – key findings

The team’s ability to act in concert without the need of explicit communication is referred to as coordination (MacMillan et al., 2004). Three critical elements to this is backup behaviour, mutual performance monitoring and adaptability.

Backup behaviour was found essential in pilotage operations, particularly in dynamic situations where the demand would rapidly shift from low to high. Backup behaviour involves the team’s ability to recognise a problem in the workload distribution, but also being able to shift the work responsibilities accordingly. This entails that a captain “*must* *be ready to engage if necessary*", as stated by one of the captains. A relevant example could be that a sudden and heavy fog surrounds the ship when sailing. This fog could be overwhelming for the pilot's workload in maintaining an overview of the situation; thus, the captain should then assist the pilot by, e.g., taking watch over the radar.

Furthermore, backup behaviour was also found to be important for agents external to the pilot-bridge team. The interviewee said, for instance, that tugs function as invaluable support during berthing operations and that they could be used for backup assistance if the thrusters were unreliable: "*If we have tugs, they can work as a backup in case the thrusters do not help*" – interview with a pilot. In sum, backup behaviour was considered an essential subfactor of coordination on the ship's bridge during pilotage and 20 statements was derived from the analysis.

Another aspect discussed was mutual performance monitoring, including the need to provide feedback to each other's performance. The feedback aims to improve the efficiency and safety of the operation if an agent sees improper performance by a team member. Furthermore, it was stressed that the person giving feedback should deliver the message in a manner that complements the member's feedback schemata. The following quote from one of the captains emphasises this issue: "*It is important to observe each other's actions and nicely provide feedback if someone does a mistake or something*". As insinuated, exact knowledge of how feedback should be delivered is difficult to conceptualise due to differences in culture, language, and individual differences.

It was also found to be imperative that lower-ranking officers know they can give feedback – even challenge – the commands of senior officers. One situation that was accentuated in connection to this was the issue of tunnel vision: i.e., the loss of peripheral vision due to, for instance, high workload. A low-ranking officer could help the pilot to consider an incoming ship that might have not been recognised when the pilot initially gave a command. This flexibility is important for the team coordination. However, it was also acknowledged that this was influenced by cultural differences.

Lastly, the concept of adaptation was discussed by the interviewees, but only to a lesser extent. It received 6 statements, in which most was referring to the significance of being vigilant during the operation. For instance, a captain stated that: "*the pilot rarely tells us that he needs help, so we (the crew) must be cautious and see if the environment changes and that something must be done*". Nonetheless, considering that adaptability was not comprehensively addressed in the interviews, is an important discussion as well.

An essential element of adaptation is the team's ability to conjointly adapt to changes in the environment (Salas et al., 2005). There was less emphasis in the interviews on the team's role in adapting to the environment, and more attention to the individual's responsibility, e.g., a pilot stated that: "*when I (the pilot) board the vessel, I start assessing how I should approach the crew and the situation*".

While team research generally suggest that teams must conjointly adjust their strategies based on environmental cues, teams in complex pilotage operations normally get less time to learn each other's way of working and could perhaps be considered as a swift starting action team (McKinney Jr et al., 2004).

This characteristic of piloting teams suggests that it would be troubling to expect that these teams can conjointly adapt to key changes in the operation. It could also contribute to explain why adaptation received fewer statements in the content analysis. Please see Table 15 below for example statements.

Table 15: Example statements coded as coordination.

Interview	Statement	Subfactor
Pilot	*"It may be that I call colleagues on the phone to discuss operational challenges, perhaps have they been there before. Maybe they know something more about the weather, fog, the specific vessel, or any potential rules and regulation to know of."*	Backup behaviour
Captain	*"We (the crew) must be ready to take initiative and help him (the pilot) if he needs extra hands."*	Backup behaviour
Pilot	*"I like to monitor the helm's actions to ensure that he does what I say. If I say starboard 10 then it's not portside."*	Mutual performance monitoring
Captain	*"Another task for captain is to observe and monitor the situation. I need to be able to help out the situation in a second's notice."*	Mutual performance monitoring

5.3.5 Shared mental models – key findings

It was found to be generally problematic for the pilot-bridge team to rely on having shared mental models. The interviewees expressed themselves scarcely concerning this issue for three of the four subfactors: 1) shared model of how to use the tools and equipment, two statements, 2) how to best communicate within the team, eight statements, and 3) having a shared understanding of each other's competencies, seven statements. Conversely, the shared knowledge of how a task should be carried out was well accentuated in the material (seventeen statements).

Shared mental models are considered central for effective coordination (Converse et al., 1993). To have established a set of cognitive models of how to collaborate, often support coordinative behaviour, such as adaptability. These schemata enable the operators to conjointly expect and adapt to future actions (Converse et al., 1993). However, the development of such schemata takes time and is consequently often absent in swift starting action teams, which are composed of trained professionals with no prior knowledge of others on the team (McKinney Jr et al., 2004).

Thus, teams in pilotage operations cannot expect that team members have certain shared mental models in the same manner that other tactical teams can, particularly teams that have been training together for an extended period, such as a military squad.

However, a specific aspect of shared mental models was frequently discussed by the interviewees: the concept of having shared models for how to carry out their tasks. This finding implies that the seafarers could rely on having a shared model for how essential tasks are carried out. Perhaps the STCW requirements, which standardises the training and education of seafarers, might be contributing to seafarers having established a shared mental model for task execution. Seafarers do receives similar training through various standardised education programs around the globe (Sampson, 2004). This training regime can perhaps explain why they can mutually rely on each-others task execution, while not sharing mental models for how relationships and communication should develop. Please see Table 16 below for example statements.

Table 16: Example statements coded as shared mental models.

Interview	Statement	Subfactor
Pilot	*"We (the pilots) want to clarify roles early, such as who has the handles during berthing. Either I do it or the captain does it."*	Shared mental model: job
Captain	*"A challenge with pilotage is that a new person will enter the bridge team and must collaborate well fast and effectively. But this can be difficult when you don't know each other's way of performing the tasks."*	Shared mental model: job

In sum, the following seven subfactors of teamwork was, based on this analysis, integrated to the CAPA-tool: 1) information exchange, 2) phraseology, 3) backup behaviour, 4) mutual performance monitoring, 5) team orientation, 6) mutual trust and 7) the shared job mental model.

5.4 Article 4 – findings

The CAPA-tool is reported in Article 4 in this doctoral research. Then, the tool's validity and reliability were examined. The current section presents the tool and then the findings from the doctoral study examining its validity and reliability.

5.4.1 The computer assisted performance assessment tool (CAPA-tool)

The computer assisted performance assessment tool, coined CAPA-tool in this research, is presented with respect to its qualitative and quantitative features. The qualitative features correspond to the theoretical underpinnings supporting both the empirical and performance indicators that were used to structure the assessment hierarchy. This hierarchy also includes the connections between the performance indicators. The quantitative features, then, refer to the weight of the factors, illustrated by the arcs (i.e., the arrows). The strength can be defined by statistics, probabilities, and expert ratings through tools such as the AHP.

5.4.1.1 Qualitative features of the CAPA-tool

The CAPA-tool consists of two main dimensions (technical and teamwork) and two auxiliary dimensions (boarding the pilot and berthing the vessel). The instrument also has an external dimension with the purpose of correcting the performance score based on the external complexities of the operation. In total, five dimensions constitutes the qualitative features of the tool.

The required technical competency has a long history and is widely covered with regards to performance assessment. Protracted work has previously been carried out concerning the identification of empirical indicators that address seafarer's technical competencies; for instance, in professional assessment protocols used in full-scale simulator assessment and dedicated research on pilotage operations. In the current doctoral research, it was considered necessary to respect the protracted work employed over the years. For achieving this, three approaches were used: workshops and task decomposition with subject-matter-experts, adaptations of existing tools for assessing technical competencies (e.g., the already mentioned assessment protocols), and lastly, research on competency requirements in pilotage operation, e.g., Norros (2004). The technical indicators, then, comprise information from all the three sources. In Table 17 below, three examples of technical performance indicators and their respective empirical indicators (i.e., the questions presented for the assessors) are given.

Table 17: Three examples of technical performance indicators and one example of their respective empirical indicators.

Performance indicator	Description of performance indicator	Example of empirical indicator
Ship-handling	Concerns the seafarer's handling of the ship.	Has the bridge team mostly used speed correctly?
Ship position overview	Encompasses the seafarer's overview of the ship's relative and absolute position.	Is proper lookout ensured?
Route planning	Involves the establishment of a sailing plan for the remaining voyage.	Is the pilot advice properly accounted for?
Ship-handling	Check the use of various mechanisms for ship transversing the water, like propulsion and rudder.	Can you observe the correct use of the throttle for the most part of the operation?

The factors in the CAPA-tool representing the teamwork competencies in pilotage operations are underpinned by extensive research on generic and maritime teamwork requirements. The four core teamwork factors and their conceptualisations were applied to the case of pilotage operation in Article 3. Based on the findings in this article, the teamwork performance indicators for the CAPA-tool were defined and operationalised as empirical indicators that the assessors would use for evaluating the teamwork competency of the pilot-bridge team (See Table 18 for three examples).

Table 18: Three examples of teamwork performance indicators and one of their respective empirical indicators.

Performance indicator	Description of performance indicator	Example of empirical indicator
Information exchange	Refers to what information is delivered between the sender and receiver.	Is the correct information shared in a timely matter?
Phraseology	Refers to how the information is delivered between sender and receiver.	Is the proper terminology used?
Mutual performance monitoring	The ability to develop common understandings of the team environment and apply appropriate task strategies to monitor teammate performance.	Is the team clear about their responsibilities?

The auxiliary dimensions are also essential aspects of the operation. Boarding the pilot safely and efficiently requires careful planning and execution to ensure that the pilot is safely installed on the bridge. Also, berthing the vessel pertains the process of getting the ship safely alongside the port. In this process, it is crucial that the ship's speed and course is correctly adjusted, as well as the use of sideway propulsion (e.g., using thrusters or tugs). Table 19 provides examples of performance indicators for boarding the pilot, berthing the vessel and external factors.

Table 19: Example of performance indicators for boarding the pilot, berthing the vessel and external factors.

Performance indicator	Description of performance indicator	Example of empirical indicator
Pilot boat adjustments	This regards proper communication with the pilot boat.	Is correct side for going alongside main ship given to the pilot boat?
Inform pilot of ship condition	Concerns assuring that the pilot is properly informed about the ship's condition.	Is a proper discussion between the bridge team and the pilot ensured?
Ship and course	Regards setting proper speed and course for the berthing operation.	Is the speed and course proper for berthing?
Berth positioning	Concerns the preparation of the ship before berthing is commenced.	Is the bridge informed of the port requirements?
Weather impact	This allows the rater to determine if the weather has a profound impact on operational performance.	Is the weather impacting normal operation?
Traffic density	This allows the rater to determine the traffic density.	What is the traffic density: little, normal, or heavy traffic?

The external factors in the CAPA-tool are time pressure, traffic, and weather (including hydrodynamic forces). These factors are considered as being out-of-control for the pilot-bridge team and serve as adjusting factors to the performance score. That is, a higher impact of external factors provides a weighed boost towards the final performance score. There is certainly a plethora of other external factors to include, however, these factors were considered to have a major role in the execution of a pilotage operation and was thus included in the CAPA-tool.

An extraction of four indicators are given in Figure 14 immediately below. In this figure, the empirical indicators are emphasised to show how the assessors' inputs are fed into the tool's higher hierarchical levels. Figure 14 is extracted from the bottom of Figure 15 on the next page.

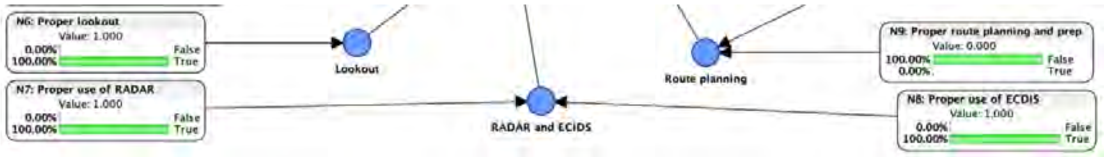


Figure 14: Extraction of four indicators (bottom part of Figure 15 on the next page). N6-8 are empirical indicators. The "N" letter shows that the indicators are part of the navigation dimension. In this example, we see that three out of the four empirical indicators are evaluated as "true".

The entire network is graphed in Figure 15 below. The two main and the two auxiliary dimensions (comprising the operational performance score), in addition to the external factors' impact are emphasised using a measurement symbol that gives a visual cue of the scoring. The final performance score, with a red highlight in Figure 15 on the next page, was calculated based on the operational performance adjusted by the raters' interpretations of the external impact. The network was created and calculated using BayesiaLab™ 8.1 software.

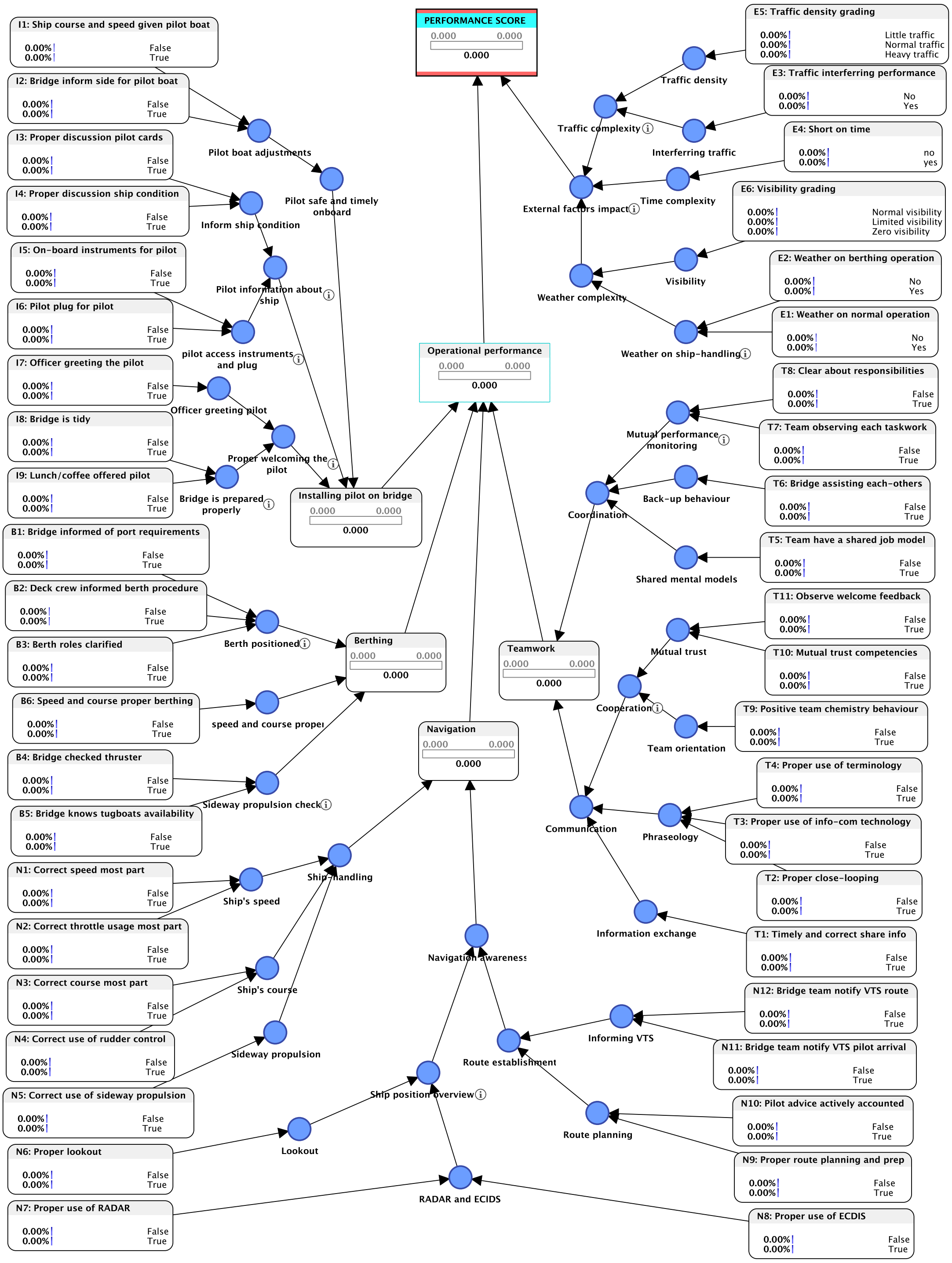


Figure 15: The CAPA-tool.

5.4.1.2 Quantitative features of the assessment tool

There are two types of probabilities that must be quantified in a BN: prior probabilities of independent variables and conditional probabilities of the dependent variables. Translating this to the assessment tool, the prior probabilities refer to the empirical indicators' measures; whereas the conditional probabilities refer to the weight assigned the relationships between the performance indicators or between the performance indicators and the empirical indicators (the probability of performance given the priors, i.e., the empirical indicators). The prior probabilities are collected by the raters when observing and evaluating the operation, while the conditional probabilities are derived theoretically or statistically, or calculated and defined based on AHP. The conditional probabilities are used to specify the probability for each dependent variable (e.g., to score the information exchange performance indicator), and we do this for every state of its directly dependable parent variable (e.g., to further give a score to the communication key performance indicator).

Worth mentioning, the conditional probabilities can also be defined using statistical data which may help tune the CAPA-tool in future iterations, which enables the instrument to be expansive as more training and assessment data become available with time and use.

Relationship between the parent and child nodes. The empirical indicators are the parent nodes and the higher order performance indicators are the respective child nodes. The empirical indicators serve as the checkmarks that are evaluated by the raters. All main empirical indicators are binary, either yes or no. However, there are also two special indicators for grading visibility and traffic density that have three response alternatives. The main empirical indicators are formulated as yes or no with the purpose of making interpretation less ambiguous: either the condition is accepted or disregarded.

However, this dichotomy can make it difficult for the assessor to make a judgement as the operation may have different situations rendering a condition accepted at one part of the voyage and unaccepted at a later part. At the same time, this places the responsibility of correctly evaluating whether an empirical indicator is satisfactory or not, for the particular scenario that is being carried out, on the present and expert assessor.

The empirical indicator input from the raters serves as a prior probability for their respective child node. The child nodes of the empirical indicators are the performance indicator. The conditional probability tables for the child nodes (i.e., the performance indicators) are weighted based on the AHP inputs. Please see Table 20 below for an example of a performance indicator's weighting scheme, and then Table 21 for its translation into the conditional probability table.

The respective weights, then, correspond to the performance indicator when it is the only *true* condition. For events in which two of the three conditions are evaluated as true, then both weights are added together and serve as an aggregated weight. Thus, for conditions where all inputs are true, a performance score of 100.0 is certain, and 0 for the opposite condition.

Table 20: Example of the weighting scheme.

Berth positioning:	B1: Bridge informed of port requirements B2: Deck crew informed of berthing procedure B3: Berth roles clarified	B1 38.5	B2 24.0	B3 37.4

Table 21: The conditional probability table for berth positioning. Please note that the B1-3 corresponds to the coding in Table 20 and the main Figure 15 above.

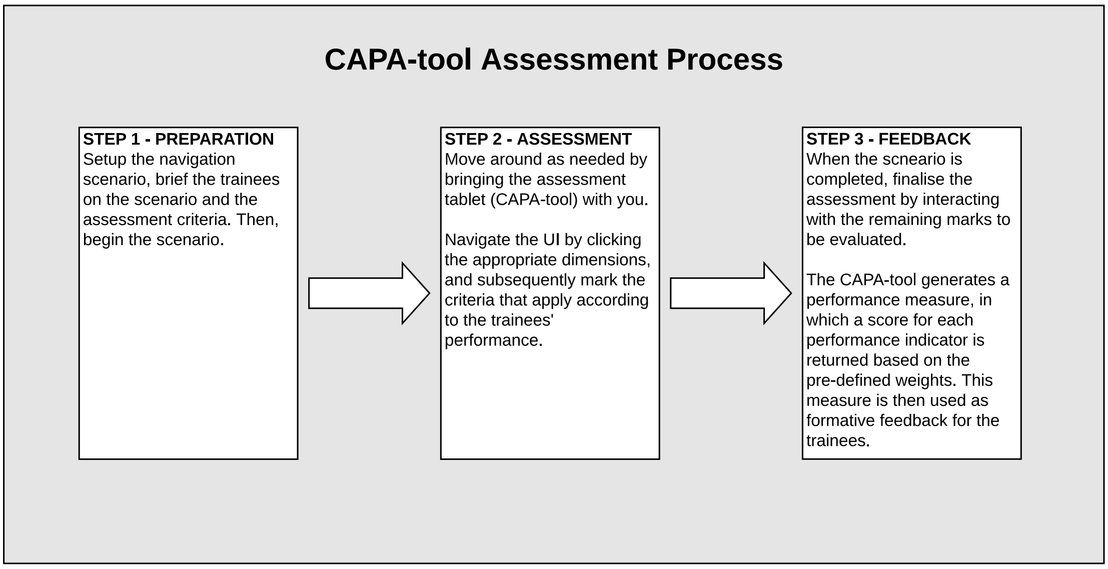
B1	B2	B3	False	True
False	False	False	100.0	0.0
		True	62.6	37.4
	True	False	76.0	24.0
		True	38.6	61.4
True	False	False	61.5	38.5
		True	24.1	75.9
	True	False	37.5	62.5
		True	0.0	100.0

5.4.1.3 An example of practical application of the CAPA-tool

The CAPA-tool, as mentioned, was developed to improve the objectivity in simulator assessment. Earlier attempts (and not reported) at objective assessments have provided necessary information concerning the need for flexibility and ease-of-use, which is also supported in Conole and Warburton (2005). Challenges with earlier attempts at objective performance assessment have disregarded the extensive resources that are required to tailor the tool for each scenario. Rigid objective operationalisations, e.g., technical simulator parameters such as cross-track error had to be re-calibrated for every scenario, and even then, could have questionable validity – what if the trainee chose a different, but equally safe and efficient turn? The learnings from earlier attempts were highly appreciated in the current doctoral research and for the conceptual design of the CAPA-tool: learning the importance of elements such as flexibility, having less rigidity as well as being swift and easy to use.

The CAPA-tool can be used for simulator assessment in its current form, although it is perhaps unnecessary cumbersome to use. The checklists that were used in the experiment reported in Article 4 had to be manually analysed by inputting the data into a computerised tool that calculates the performance score (i.e., it had to be run in the BayesiaLab™ 8.1 software). However, this process can be and is intended to be further automated by using interactive computer application to input the data, in contrast to using a manual checklist, which was necessary for the current experiment. This application, then, can forward the information and immediately return the performance score for the evaluator and the trainees.

The computer application can run on a tablet, which also enables the assessor to be mobile when assessing using the CAPA-tool. For instance, there may be a need to move between the instructor station and the simulated bridge. Meanwhile moving around, the assessor can check-off the relevant categories that he or she sees fit during the simulator exercise. The CAPA-tool evaluates the input and weight the factors following an objective (or intersubjective) agreement among experts and returns a formative performance score for the trainees. Then, after the exercise, a report is automatically generated that breaks down the areas and highlights the ones that requires attention. This report may be a foundation for debriefing the maritime exercise and to show the need for further training. Please see Figure 16 for an example of how the CAPA-tool could be used.



CAPA-tool Assessment Process

STEP 1 - PREPARATION
Setup the navigation scenario, brief the trainees on the scenario and the assessment criteria. Then, begin the scenario.

STEP 2 - ASSESSMENT
Move around as needed by bringing the assessment tablet (CAPA-tool) with you.

Navigate the UI by clicking the appropriate dimensions, and subsequently mark the criteria that apply according to the trainees' performance.

STEP 3 - FEEDBACK
When the scneario is completed, finalise the assessment by interacting with the remaining marks to be evaluated.

The CAPA-tool generates a performance measure, in which a score for each performance indicator is returned based on the pre-defined weights. This measure is then used as formative feedback for the trainees.

Figure 16: Example of the CAPA-tool assessment process. Please note that the assessment tablet is not within the scope of this doctoral thesis (a manual checklist was used as a substitute in the experiment reported in Article 4).

5.4.2 Examination of the CAPA-tool’s reliability and validity

The reliability and validity of the assessment tool were examined in Article 4. A maritime simulator scenario was developed for expert raters to assess. The expert raters were separated in an experiment and a control condition. The experiment condition used the CAPA-tool and the control condition used conventual assessment methods.

Interrater reliability, absolute reliability, construct validity, content validity and criterion validity were addressed in Article 4. However, the tool’s content validity was already theoretically supported by the previous research on the topic (particularly for the teamwork and technical dimensions). The criterion (predictive) validity, however, will be investigated as future performance data are gathered by correlating the performance score of the tool with true performance data in the future. Interrater reliability, absolute reliability and construct validity were covered by Article 4’s research questions.

5.4.2.1 Interrater reliability of the assessment tool

The interrater reliability was calculated using Krippendorff's alpha. There were nine raters for the analysis, 47 cases and a total of 409 decisions to evaluate. It was the same nine participants that were placed in the experiment group. The interrater reliability was determined as fair (α = .31) after excision of the chance of having similar answers. This result is not considered as satisfactory as it implies too much variation among the raters when following Krippendorff's strict interpretation of interrater reliability estimates, in which 0.667 is regarded as the absolute minimum (Krippendorff, 2018). However, at the same time, an α of .31 could suggest a step in the right direction considering the complexity associated with assessing maritime pilotage operations.

5.4.2.2 Absolute reliability of the assessment tool

The absolute reliability of the assessment tool's final score was assessed by referring to the CV for each of the experimental condition. The experiment group outperformed the control group with regards to the reliability of the technical score. However, neither groups returned viable reliability estimates for the teamwork score or the total score. See Table 22 below for the total mean score and CV for both the experiment and control group.

Table 22: Total mean score and CV for the experiment- and control group.

	Total		Technical		Teamwork	
	Mean score	CV	Mean score	CV	Mean score	CV
Experiment	42.44	31 %	61.59	14 %	38.59	62 %
Control	47.06	36 %	43.25	50 %	42.88	42 %

Furthermore, interpreting and extrapolating the meaning of the CV is difficult. There is an arbitrary convention of desiring a CV less than 10 %; however, this must be considered on a discipline basis. For the current application, 14 % dispersion, which was achieved when rating technical performance, is argued to represent evidence of a reliable measure of technical performance in this current application.

5.4.2.3 Validity of the assessment tool

In the experiment in Article 4, the mean deviation from the gold standard was 32.56 (the group using the CAPA-tool) and 29.19 (group using the conventional assessment). The mean deviation was significantly different from the gold standard for both the CAPA-tool ($M = 32.56$, $SD = 13.32$, $t(7) = -9.01$, $p < .05$) and the conventional tool ($M = 29.19$, $SD = 14.25$, $t(7)$, $p < .05$). This finding fails to support and find evidence for the criterion validity of the CAPA-tool (nor the conventional tool).

Examining the reliability and validity of the assessment tool was the final stage of this doctoral research. To sum, the CAPA-tool shows promising reliability estimates concerning the use for assessing technical performance. It also shows strengths regarding relevant features for assessment in training and education purposes, such as providing detailed and intersubjective feedback for the trainees and students. At the same time, more studies should be carried out to further evaluate the validity of the assessment tool.

The information from this study suggests several paths and areas of improvement moving forward. These opportunities will be further discussed in the next chapter.

6 Discussion

This chapter extends and discuss the findings in this doctoral research by connecting it to the greater scientific body of knowledge on training and assessment, and to practitioners in the maritime domain (section 6.1 and 6.2). Then, a discussion on the tool’s validity and reliability, as well as methodological reflections and considerations for the various stages of the doctoral research (section 6.3 and 6.4). The chapter concludes with an example of applying the CAPA-tool in assessment before a presentation of the outlook of future research directions and recommendations to practitioners (section 6.5 and 6.6).

6.1 Performance assessment in maritime education and training

Recent literature reviews report a need in the maritime industry to focus on the use of bridge simulators (Sellberg, 2017) and on the methods used when assessment methods are developed (Ernstsen & Nazir, 2018a). This need is aggravated by extensive research on the importance of effective training and assessment. Article 1 in the current doctorate research examined the development of assessment methods across four maritime domains: port logistics, shipping and navigation, safety, and environment, and suggested more attention to evaluation procedures in shipping and navigation by developing structured methods with clear and articulated assessment criteria.

Unstructured assessment with the use of implicit assessment criteria are prone to high degrees of subjectivity (Moorthy et al., 2003). The evaluators are prone to personal biases; such as serial positioning effects (Murdock Jr, 1962), halo effects (Nisbett & Wilson, 1977b) and recognition-primed inferences (T. D. Wilson & Brekke, 1994), which makes it difficult to observe and assess situations objectively.

The lack of structure in simulator assessment has been found to be detrimental for maritime safety (Gekara et al., 2011), in which an unfortunate assessment framework can orient the learning environment away from acquiring the necessary skills and knowledge, towards having trainees that rather focus on how to perform better on the competency tests (Emad & Roth, 2008).

Section 3.1 introduced the importance of proper feedback in training and assessment, conceptualised as the information given by a trainer regarding aspects of the trainee's performance (Hattie & Timperley, 2007). In this, the feedback must provide information that minimises the gap between what is understood and what is aimed to be understood (Ramaprasad, 1983; Sadler, 1989). A key element of proper feedback, as argued in Taras (2005), is the use of clear and articulated assessment criteria - in contrast to the implicit and subjective criteria residing within the evaluator (and thus more prone to the personal biases). Only by developing and using formulated criteria can the assessor give proper feedback to the trainer, re-assess the scenario using another evaluator, and compare the performance across trainees for refining and improving the training and assessment program.

Furthermore, the effect of teamwork training in the maritime has been questioned. It has been discussed that the training programmes have not been tailored according to the specific needs of various maritime operations (O’Connor, 2011; Salas, Burke, Fowlkes, & Priest, 2004). However, in light of the findings in this doctoral research, it could also be argued that the lack of teamwork training effect is due imprecise measurement of teamwork performance, and especially for particular operations such as a pilotage operation. It is therefore suggested to provide the trainees and students with reliable and valid feedback in the performance assessment of pilotage operations in full-scale simulators.

6.2 Assessing pilotage operations in full-scale simulators

Assessment is a part of being human. We constantly, mostly implicit, assess the world around us by using empirical data to refine our behaviour and beliefs. This process is an essential part for us to learn, grow and adapt to the feedback we give and receive from assessment. Humans have developed efficient cognitive mechanisms to carry out these assessments, such as the representative, availability and anchoring heuristics (Tversky & Kahneman, 1974). These mechanisms are invaluable aspects of human cognition where most situations call for a quick and approximate decision, and not requiring a precise decision (Stanovich & Toplak, 2012). However, if an environmental stimulus does present itself, a slower and more precise cognitive process can be triggered (Evans, 2008).

While our everyday assessment does not need to be systematic and objective and can benefit from being fast and approximate, educational and professional assessment must be. Educational and professional assessment refers to a systematic process of documenting the use of empirical data on a trainee's or student's competency. This systematic process establishes standardised methods for assessment that experts and research community can collectively agree and adhere to, thus making it more objective. Striving for objectivity in assessment is particularly critical in ensuring that workers and operators possess the competency necessary to safely and efficiently carry out their job.

However, by assessing simple tasks, for instance turning on a computer, one can count the power button's hit-ratio as an objective assessment. Clearly, the development of objective assessment methods for complex operations is more complicated than merely counting hit-ratios, and as operations turn even more complex, obtaining an objective assessment method can become highly resource demanding. As introduced in chapter 2, complex operations consist of a multitude of interdependent factors that the operators must constantly consider. This complexity contributes to, that in many instances, the assessment of complex operational performance remains subjective. However, for high-risk operations, it might be considered worthwhile to invest the necessary resources for a dedicated assessment framework.

The CAPA-tool's flexibility is an essential characteristic considering the high cost associated with the development of the instrument. One way, as previously discussed, is to have humans in the loop. As an example, the correct use of the RADAR settings might change for a particular exercise (e.g., due to weather and traffic variations) and might therefore be rated according to the designed scenario. However, the weight attributed to the RADAR competency remains intersubjective; that is, a fixed and agreed upon measure set by several experts in conjunction about how important the skill and knowledge of using the RADAR is. This characteristic of being malleable for different scenarios was necessary for it to be practically and user friendly.

The assessment tool developed in the current research demanded attention and input from a plethora of different subject matter experts and had to accommodate for a wide range of competencies. As an example; in stage 2 of this research, it was necessary to collaborate with computer engineers to develop the conceptual assessment framework; whereas in Stage 3, it was necessary to use expert navigators for the technical competencies, expert pilots for the pilotage factors, and expert teamwork researchers for the bridge resource management factors. Article 2 reports a conceptual assessment framework in light of the research carried out in stage 2. The concept was developed to reduce the subjective impact, while being flexible, and was done in close collaboration with a computer engineer to ensure proper use of AHP and BN.

Stage 3, then, concerned expanding the qualitative features of the tool. The technical features in which the tool consists of have been widely researched; thus, the scope in stage 3 focused on establishing the teamwork requirements that best fit the case of pilotage operations. In this stage, it was found that the expert seafarers – pilots and captains – were acknowledging the need for communication, coordination, cooperation and shared mental models, also in line with the prevalent theories on teamwork, but with one exception: The teamwork factors that requires a sustained working relationship were found to be less important in pilotage operations. While they would be valuable for performance, it would not be enough time to effectively develop these factors in such swift operations and should perhaps receive dedicated attention.

6.3 Validity and reliability of the CAPA-tool

Article 4 studied the validity and reliability of the performance assessment using the CAPA-tool and using conventional assessment methods. The study found evidence of acceptable reliability when assessing the technical elements of the pilotage operation, but only when the raters were using the proposed CAPA-tool. The conventual methods for assessment, however, returned a too high dispersion among the raters. This finding is interesting concerning the need for feedback as the assessment method must be reliable for any feedback to be meaningful (Kimberlin & Winterstein, 2008). By using the proposed assessment tool, the evaluator can provide the trainee with feedback on their technical performance that is supported by several experts (and not only the subjective performance interpretation of the current evaluator).

The study reported in the fourth article, however, did not find evidence of reliability for either the teamwork component of the assessment tool or for teamwork when employing conventual methods: both groups of raters returned too high dispersion on their scoring. This finding implies a need for utmost care in returning formative feedback concerning teamwork performance, which is underpinned by other research on BRM, e.g., O'Connor (2011) and Salas, Wilson, Burke and Wightman (2006). There is still a lack of an objective agreement on how to reliably assess teamwork competencies.

6.3.1 Reliability and validity in teamwork

Understanding teamwork is undeniably challenging. Scientists and practitioners have conducted a plethora of research and development concerning teamwork in recent years. Early, the non-technical skills taxonomy presented a framework for assessing teamwork, which was a critical element in the development of crew and bridge resource management training and assessment programmes (Flin et al., 2003). From a general perspective, Salas et al., (2005) frames a selection of critical factors for team performance, while Rafferty et al., (2010) advances the general understanding of teamwork in complex operations by identifying four core teamwork factors (communication, coordination, cooperation and shared mental models) and applying them to the case of military fratricide.

However, there are other perspectives to team performance that could serve viable for the reliable and valid assessment of teamwork as well. For instance, distributed cognition in which the entire system is incorporated into the assessment has certain advantages (Hutchins, 1995), and is a perspective that should be considered for when the performance assessment aims to capture performance "in the wild" (to borrow the words from Edwin Hutchins), i.e., outside of the controlled simulator environment. This perspective takes a broad view and incorporates the effect of culture on the individual as well. In which case, distributed cognition could be a critical element to further increase the CAPA-tools ecological validity in the case that the assessment tool was adapted to outside of the simulator environment.

6.3.2 Teamwork in the current research

The deductive approach in Article 3 was central and helpful for providing proper nomenclature and explanations for examining the core teamwork factors for the case of pilotage operations. However, achieving cooperation in maritime operations is difficult. One reason is that teams that are operating on the ship bridge, especially during pilotage, includes a multitude of different nationalities and cultural backgrounds. In addition, achieving a similar safety climate has been shown to be difficult even across companies (Mallam, Ernstsen, & Nazir, to be published). This multiculturalism, in light of the discussion above, may inhibit the development of team cooperation because of the lower levels of homogenous attitudes and preferences. This difficulty could be argued to be aggravated in pilotage operation, in which the pilot-bridge team must perform without sustained working relationships and the development of coordinative mechanisms for swift starting action teams is difficult (McKinney Jr et al., 2004).

On the other hand, skills and ability requirements within a ship bridge team is standardised according to the STCW regulations (IMO, 2011), which may help to ensure skill and ability heterogeneity within the bridge team. This standardisation is certainly be beneficial for some elements of teamwork, but perhaps the standard could be strengthened to focus on other (relevant) aspects of teamwork as well. Especially considering that the maritime environment comprises vast variety of equipment, tasks, roles, and team members. It is imperative for a safe and efficient voyage that there exists a shared understanding among the crew. However, while the bridge team has plenty of time to form and develop shared mental models, a complex pilotage operation relies on close interactions with a temporary pilot and in many instances, this gives the team insufficient time to develop the efficient shared mental models.

Furthermore, Goodwin (1994) discusses the concept of intersubjective professionalism concerning how practitioners develop a common way of perceiving a situation. In a highly globalised and multicultural industry such as the maritime domain, achieving intersubjective professionalism can be difficult. In fact, a recent study has examined the importance of focusing on demonstrating this professionalism for the students and trainees (Sellberg & Lundin, 2017). Furthermore, Taylor (1998) points out that "good seamanship" is highly situation-dependent, and that the steps necessary for certain tasks are socially defined. These concepts emphasise the importance of developing shared mental models and professional intersubjectivity. Thus, in addition to demonstrating professional intersubjectivity, an introduction of these training criteria globally (e.g., through IMO) could perhaps lead to higher sharing of the remaining dimensions of mental models on the bridge, particularly relevant for pilotage operations.

Besides, factors necessary for team performance could have been investigated from an inductive standpoint instead. This approach would have entailed a need to develop teamwork factors for pilotage operations, rather than adapting current frameworks. While developing teamwork factors are resource intensive, this approach could perhaps have provided a more accurate representation of the teamwork factors in pilotage navigation; thus, returning a more reliable (and valid) teamwork measurement.

However; ultimately, the vast amount of teamwork research over the past decades suggested the use of a deductive content analysis based on the already established teamwork research.

6.4 Methodological discussion

6.4.1 Reflections on the research methodology

The research progress and the project stages are conveyed by considering the appended articles in sequential order. The findings generated at each stage, reported in each article, is directly or indirectly contributing and transferred to subsequent stages. However, this transferring of knowledge also suggests a careful reflection of the methodological limitations associated for each of the conducted studies. Therefore, section 6.4 is dedicated to discussing relevant shortcomings with the doctoral research project's data collection and data analysis. But first, a discussion on the epistemology.

6.4.1.1 What kind of knowledge did the researcher aim to produce

The research aimed to reduce the subjective impact in maritime full-scale simulator assessment, which called for an applied research approach. A central challenge, as is seen in the methodological procedures presented in Section 4.3, was the population size. The lack of participants implied that a combination of research approaches was necessary to effectively explore and answer the research questions.

The experiment reported in Article 4 was designed to investigate the CAPA-tool's reliability despite the specialised sample. Ideally, the experiment would statistically assess both the within and between rater agreement, where either the intraclass-correlation coefficient or the Cronbach's alpha could provide more robust measures on reliability than the CV. However, due to the vast amount of resources required to run a full-scale pilotage scenario, this compromise was required. Still, enough expert navigator assessors were available to statistically assess the absolute variability between the two conditions.

Furthermore, another crucial design discussion regarding methodological design and data collection revolved around whether to employ a within or between experiment, where sample size was a main argument for the within experiment. This approach could give more statistical power as the final number would be 16 instead of 8. However, the learning effect was considered to have a substantial impact on the data's (internal) validity. Even by changing which participants that would first employ the CAPA-tool and first use the conventional methods; only the data from their respective first run could be considered unbiased. And perhaps more important: this approach would request four hours of the participant's time, instead of two hours with the between-subject design.

Lastly, investigating the construct validity was also compromised by the lack of participants, as discussed in Article 4 and in this thesis introduction. However, the combination of methods, mixed methods, carried out throughout the doctoral research enabled the researcher to explore the research questions from different perspectives and angles. This approach, however, asserts that the underlying research and the methods employed are transparently and comprehensively reported throughout the research (Chapters 3 and 4), as well as proper descriptions of the background and context in which the research is conducted (Chapter 2).

In light of the epistemological debate, the next section discusses limitations of the research data collection and data analysis more specifically. There are plenty of caveats and pitfalls associated with collecting empirical data. All methods have limitations; also, practical constraints such as time, resources, and availability of participants can generate ever more limitations to the research, as discussed above. It is important to acknowledge and consider this when evaluating the findings and its implications of a particular research. In the following sections, limitations in the current doctoral research concerning data collection, data analysis and for the CAPA-tool itself are discussed.

6.4.2 Data collection limitations

6.4.2.1 Systematic literature review limitations

The Boolean key string that was chosen for the literature review has implications on which papers are returned from the search. However, careful considerations were made. For instance, the word “maritime” was coded as “marine” and “maritime” to ensure that researchers using either terminology would be reflected in the search. Another limitation was the availability of papers concerning language, databases and journals. The search and review of literature was aimed for English research papers. This precludes research published in other major languages, such as Russian, Spanish, Portuguese, and Chinese research. The same limitation pertains the databases and journals available in the current literature review. As a countermeasure, the researcher ensured transparency by comprehensibly list the databases relevant in the review.

Finally, limitations concerning the exclusion criteria must be me addressed also. The exclusion of papers is a necessary step in literature review as the key-string can easily return non-relevant papers. In the current study, the keywords “marine” and “performance indicator” returned papers related to marine biology studies, which was not relevant for this literature review. Articulated exclusion criteria were established; however, it also influenced which research papers that remained for the analysis. Therefore, it was necessary to be transparent in the exclusion process also, in which the number of papers removed in each step of the process was listed.

6.4.2.2 Semi-structured interview limitations

An overarching critique and limitation of semi-structured interviews, according to Potter and Hepburn (2005), is how researchers neglect contextual features of the interview and consequently take the data at face-value. Even if the interviews were carefully prepared and planned in the current doctoral research, contextual and personal features of the interview are influencing the material. The reader must consider this limitation when addressing and evaluating the findings. Furthermore, for the interview reported in Article 3, the researcher was conscious of proper interview techniques for ensuring more meaningful data, like planning, engaging, accounting, closing and evaluating (i.e., the PEACE-model) all of the interviews, as outlined in Subsection 4.2.2. Besides, semi-structured interviews are compatible with a wide range of data analyses (Willig, 2013).

Characteristics of the interview also impact the results. The interview sample (N = 10) and length (M = 64 minutes) were adequate considering the specialisation of the research questions (i.e., navigational expertise), the limited population pool, and limited research resources available. However, critics may argue (and reasonably so) that the researcher could have achieved a broader theoretical representation with a larger sample. However, the researcher attempted to counter-measure this by purposely draft the participant to ensure a balanced representation of expertise.

6.4.2.3 Experiment limitations

The experiment is a scientific method for hypothesis testing. However, there are pitfalls that must be addressed when designing an experiment. Two different experiment approaches are the within-subject design and the between-subject design. Commonly recognised limitations exist for both. Within-subject experiments must be conscious of learning effects, e.g., a third factor makes some people faster learners, while the between-subject experiment must address various types of assignment biases (e.g., expectancy bias – the subject expects an effect and therefore unconsciously impacts the outcome). The experiment reported in Article 4 was a between-subject design.

There are three limitations in the experiment carried out in this doctoral research that need to be discussed. The first limitation concerns the experiment sampling procedure. The pool of participants was small because of strict inclusion criteria. The participants had to be expert navigators and with simulator training experience. The sampling procedure; thus, was pseudo-randomised to ensure sufficient representation in both experiment conditions. The participants were also stratified from two unique assessment facilities and only from these two facilities. Including more or different strata may have impacted the results also. A randomised sampling involved a too high probability of skewing the sample size in either condition: this may have reduced the accuracy of the sample to represent the larger population.

The second limitation pertains to the compromises made in balancing internal and external validity in connection to the complexity of the experiment procedure. As mentioned in section 4.2.2, the internal and external validity of the experiment must be carefully balanced: a more controlled design makes the testing environment more artificial, thus more difficult to generalise findings out of the laboratory setting. The participants were asked to assess an 87-minutes long navigation scenario. In this, a plethora of interdependent factors needed consideration.

For instance, it was decided to not standardise the procedure entirely, as this would have limited the generalisability of the result: This decision was connected to the flexibility design requirement of the assessment tool. However, the autonomy that the participants were given when evaluating the scenario gives less control of third factor parameters that may have impacted the results. For instance, the participants were allowed to continuously assess the scenario description given during briefing, they were sitting in slightly different assessment environments, the participants could freely attend to bathroom- and coffee needs and have their phone available (which some participants used). It is therefore difficult to rule out that any third factors may have distorted the results.

The third limitation concerns the navigation scenario that was used for assessment. The scenario was a well-known sailing route that most navigators in the area knew from personal sailing and from training and education programs at their respective facilities. However, clearly the degree of experience would be different, where certain assessors would have more experience than others. This effect can be considered an unsystematic error in most experiments; thus, statistically nullified. However, the current sample size may insufficiently rule out the error, suggesting that the result may have been impacted by the unsystematic distortion of the result. However, at the same time, the participants are trained assessors and may have evaluated the scenario irrespective of their knowledge of the sailing route.

6.4.3 Data analysis limitations

6.4.3.1 Qualitative data limitations

The qualitative data was analysed using a deductive content analysis. This method is a powerful tool for generating meaningful insight from textual data; however, it is also subject for limitations.

A recurring limitation of qualitative research in general is particularly relevant for the content analysis also; that is, its dependence on the researcher's individual skills, personal biases and idiosyncrasies. This limitation makes it difficult to demonstrate scientific rigor and elevates the importance of reflexivity (Willig, 2013), i.e., explore how the researcher's involvement influences the research (Nightingale & Cromby, 1999, p. 228). In many instances, qualitative research aims to be explorative and open up other research areas, and an extensive theoretical description of the phenomenon is the contribution.

However, in the current doctoral research, the method was used to explore the applicability of a theoretical framework (i.e., the method was top-down oriented) - that subsequently would theoretically support the assessment tool. This purpose suggested a need for keen attention to its rigor and further investigate how the findings were impacted by the researcher's individual skills, personal biases and idiosyncrasies. As a countermeasure to this limitation, a comprehensive interrater agreement testing was carried out with several raters for investigating the level of agreement among the raters (after excluding chance), as suggested in Neuendorf (2016, p. 150).

Coding is another issue that must be addressed. Content analyses relies on a coding process; however, the interpretation and operationalisation of the codes may challenge the generalisability of the content analysis. This can make it difficult to make inferences across studies. At the same time, the codes in the current doctoral research was derived from a scientific framework where the constructs are precisely and accurately formulated. This articulation of the constructs may contribute positively to the generalisability of the content analysis. Although, the topic is still abstract, so operationalised behavioural markers can still amount to variability in both observation and interpretation. This suggests, as with all qualitative research, that the reader is wary when evaluating the results.

6.4.3.2 Quantitative data limitations

The limitations connected to the analysis of quantitative data is connected to reliability and validity of how the data is collected as the calculation itself is strictly mechanical. However, there are inherent limitations of the frequentist statistical inference in research where it is hard to acquire sufficient statistical power (due to a highly specialised population). These limitations concern the subjective evaluations that must be made concerning alpha levels and the choice of the statistical test itself.

The concept of hypothesis and significance testing, and the Type I and Type II error probabilities (Neyman & Pearson, 1933) are critical in statistical inference. The thresholds set for deciding whether to support of reject the respective hypothesis are subjective conventions and can thus be misleading. Type 1 errors is when the researcher incorrectly rejects a true null hypothesis, and Type 2 errors are when the researcher fails to reject a false null hypothesis. Whichever is "worse" is determined by the context; however, the goal should be to correctly reject the hypotheses, which implies setting the correct significance (alpha) levels. These levels are mostly set by convention (e.g., *P* < 0.001, 0.05, or 0.10), where the risk of Type 1 errors rises with a less strict significance level (probability of Type 1 error = alpha level), but the risk of Type 2 error reduces (probability = beta level). An alpha level of 0.05 was determined for the statistics reported in Article 4.

Furthermore, the one sample t-test reported in Article 4 compared the sample mean against a gold standard chosen as the true mean, to evaluate if the difference were significant. A significant difference would imply that the sample were statistically too far from the golden standard. However, the main limitation when evaluating the finding in this study pertains whether the gold standard was correct, drawing an error of the third kind. At the same time, the gold standard was, given the resources available, the best approximation that could be made. But the study suggests a wary interpretation of the findings. This limitation relates to overall challenges with the overall proposed CAPA-tool.

6.5 Using the CAPA-tool for assessing performance in full-scale maritime simulators

A dedicated section is given for discussing potential challenges and opportunities concerning the CAPA-tool and the aspects that must be considered when using and further developing it.

The qualitative aspects must be considered, including the limitations discussed above. Furthermore, this also suggests (but does not prerequisite) that the user should spend time to comprehend the scientific literature prior of employing the CAPA-tool, to ensure that the depth of the empirical indicators is adequately comprehended. In light of this, a set of minimum requirements for the assessors should be developed.

The quantitative aspects must also be addressed. These aspects concern how the variables are weighted. As presented in Chapter 5, the weights are derived from the AHP. However, some factor weights did not achieve consensus (there were some disagreement among experts concerning the weights). This impacts the tool's objectivity and should be considered in subsequent iterations and tuning of the tool. This issue must also be considered when giving formative feedback based on the pertinent factors.

The generalisability of the tool is also worth to consider. The reliability and validity of the tool has been tested in a Scandinavian context. This context puts limitations on the ecological generalisability of the tool. It has been discussed earlier that a flexible tool was necessary for ensuring a wider range of applicable scenarios and situations. The human assessor of the empirical indicators makes the tool more flexible as she or he may judge differently based on circumstantial requirements. However, it has not been tested or considered whether the factor weights must be recalibrated for other types of navigation cultures, e.g., pilotage operations that takes place in Mediterranean ports.

Furthermore, the tools reliability and validity are difficult to statistically assess. Ideally, a proper exploratory- and confirmatory factor analysis would have been conducted in the development for identifying the tool's dimensionality (and to examine the construct validity). Then, the intraclass correlation coefficient for each dimension would be considered. However, this procedure would require an unattainable sample size bearing in mind the amount of resources required for running each sample. The current sampling distribution challenges the statistical rigor concerning reliability and validity estimate of the tool. However, the experiment carried out and presented in the fourth article attempts to provide reliability and validity estimates despite this statistical challenge.

An important element for the content validity was to carry out a top-down content analysis (reported in Article 3). The aim of this study was to fit and adapt teamwork research to maritime pilotage operations. Fitting and adapting teamwork research would provide higher confidence that theoretical facets of teamwork in maritime pilotage were included in the framework. The content validity is thus supported through a transparent content analysis based on 10 subject matter expert interviews with a moderate (*Cohen's kappa* = .706) interrater reliability (McHugh, 2012).

However, while this supports the tool's content validity, the criterion validity must also be addressed. The criterion validity was investigated in the experiment study reported in Article 4. It concerns whether the CAPA-tool measures what it is intended to measure. It was investigating by comparing the rater's assessments with a pre-defined gold standard. Unfortunately, the rater's assessment was statistically different from the gold standard (for both the conventional and the experimental condition). However, this result is problematised to be connected to the imprecision of the gold standard. While the researcher attempted to accurately design the scenario for the gold standard, the complexity of the operation may have inhibited a true representation of the performance. Thus, further examination of the construct validity is suggested.

Two types of reliability were examined. The first concerns the interrater reliability among the empirical indicators, the other is the absolute reliability for the summative evaluation score (i.e., the final score without any diagnosing of the results). The interrater reliability analysis returned a fair estimate of the agreement among raters (Krippendorff's nominal α = .31). As discussed in Article 4, this suggests too much variation among raters, although it may be argued as satisfactory in certain events, e.g., Landis and Koch (1977). Also discussed in the paper, actually achieving α = .31 for events with high complexity such as maritime pilotage operations may suggest that the assessment tool has a significant contribution for further improving the assessment of maritime simulator training. The absolute reliability, on the other hand, show that the assessment tool's capacity of assessing technical competence were reliable, whereas assessing teamwork was not. This finding is supported by research that underpins the difficulty associated with assessing soft skills. Furthermore, the conventional assessment method failed to suggest any reliable estimate of assessing either technical- and teamwork competency, but the CAPA-tool did find a reliable estimate for the technical dimension. This finding strengthens the CAPA-tool's contribution in further improving the field of maritime full-scale simulator assessment.

6.6 Recommendations for further research and stakeholders

Suggestions for further research and the stakeholder recommendations are connected to the central message conveyed in this research: to systematise formative simulator assessment and to make it more objective. The suggestions for future research include methods and perspectives designed to further explore and improve the applicability, reliability and validity of the CAPA-tool.

The stakeholder recommendations are grounded in the conducted research for improving the assessment reliability in a shorter, medium and longer time perspective. Although the scope of this research has been on maritime full-scale navigation simulators, suggestions for future research and stakeholder recommendations could also be relevant to other complex operations exercising simulator training as well.

6.6.1 Further research

The suggestions for further research are tied to further improving reliability and validity concerns addressed in section 6.3 and suggestions regarding further improving the utility of the assessment tool.

An objective assessment tool for complex operations is difficult to achieve. It may even be argued to be a perpetual process in which social and technological developments continuously provide new training demands and assessment criteria. It is therefore of paramount importance that research maintains it momentum for ensuring proper training and assessment programs for tomorrow's seafarers. The assessment framework presented in the current doctoral dissertation could contribute to facilitating the perpetual development of complex maritime full-scale simulator performance assessment.

Future research must further examine the reliability and validity of the CAPA-tool. The first step would include to investigate the expert consensus concerning factor weights. A potential approach would be to separate experts in cohorts for further our understanding of factors with low consensus. Perhaps patterns in sailing background can reveal that different cohort of experts prioritises differently when assessing complex pilotage operations. This information, nonetheless, will be invaluable in the development of assessment methods.

Another major step will be to strengthen the assessment of teamwork skills. This is a critical task that remains unresolved across a wide range of disciplines. For instance, a comprehensive review of assessment literature related to surgeons in the operating room failed to identify adequate behavioural marker systems for rating various teamwork skills (S Yule, Flin, Paterson-Brown, & Maran, 2006), and this struggle was sustained by subsequent empirical experiments in the same field (Steven Yule et al., 2009), as well in the maritime domain (O'Connor, 2011). Experiences from other disciplines are similar and the reliable (and valid) assessment of non-technical skills remains a challenge.

There are also suggestions relating to future research on expanding the functionalities of the CAPA-tool. One suggestion pertains user-interface studies for improving the usability of the tool. The current state is raw and not user-friendly. The user-interface research could revolve around the development of graphical applications that assessors can use to input their decisions on the various empirical indicators. This application would also enable the assessor to tailor the tool for its specific purposes, e.g., by including and excluding modules unrelated to the current scenario (e.g., removing the berthing module). This modulation will, however, have further implications on the tool's validity, which therefore must be addressed in conjunction. Lastly, a properly designed user-interface could also include and present statistics over time for the assessor to better compare results across and within students and trainees.

In a longer time-perspective, research into advancing the assessment tool for real-time assessment could generate unique opportunities to researchers (and other stakeholders). This would require, however, that the human rater is replaced by technological input mechanisms, e.g., sensors, for providing real-time objective data that an algorithm uses to infer the state of the current situation. This development will require dedicated effort in machine learning (for learning how to correctly detect empirical data) and building a strong algorithm for analysing the incoming data packages deriving from the technical sensors. An objective real-time assessment tool would be an invaluable asset for stakeholders across the maritime domain as well as other industries.

6.6.2 Stakeholders

Relevant stakeholders are MET facilities and their practitioners, such as teachers and instructors. Other stakeholders are ship-owners, local communities, pilot agencies, and ship bridge personnel. However, this discussion concerning stakeholder recommendations is focused on MET facilities and shipping companies as the assessment methods will provide more indirect effects on the other stakeholders.

MET facilities are interested in the training and educating of maritime seafarers. Providing formative and summative assessment is of principal importance to their assurance of educating seaworthy personnel. Yet, the current research found evidence that the conventional assessment methods do not return reliable estimates on the assessment of maritime navigation competencies in pilotage operations. While the assessment tool proposed in this dissertation, the CAPA-tool, lacks a user-friendly interface, its structure might be employed within a shorter time perspective for a reliable estimate of technical performance in complex navigation operations where the students and trainees are operating in conjunction with pilots. This instrument will provide the assessors and students with a formative assessment that can be used to give on-point feedback to the students. For instance, students that fails to meet the criteria for passing the course may receive specific feedback that their competency with navigational tools must be improved and which elements that should be improved.

More specifically, it is recommended that teachers and instructors in maritime navigation implement and adapt aspects of the CAPA-tool when assessing navigational performance in full-scale simulators. Although it is essential that its implementation is carried out in light of the limitations discussed herein. Moreover, by doing this they also contribute to further develop and refine the CAPA-tool, ultimately returning a more valid and reliable assessment solution.

In a longer time-perspective, maritime simulator education should be more deeply connected to assessment. A performance database could be developed that enables the course responsible to compare and contrast assessment over time. This database can be an invaluable tool for further refining and developing the course according to industry and operational demands. However, the development of a performance database requires formalised assessment methods; where the proposed assessment tool in this research may serve as such. This framework will help to connect maritime education and assessment with the ultimate goal of achieving a better prepared workforce for the maritime shipping companies.

A final recommendation, then, is to the maritime shipping companies. As the industry is perpetually developing, so must the operative personnel, but also the performance criteria used for assessing the training of these operators. A dedicated formative assessment method will provide the trainee with an overview of his or her strengths and weaknesses, enabling the operator to further refine subpar skills. At the same time, the ship-owners may receive a statistical overview of their workforce's competencies – which could, and perhaps should, be anonymised – to get strategic information concerning overall training- and hiring strategies. This information can serve as an invaluable asset in the ever-competitive market where shipping companies take a high toll on inefficient operations, incidents, and of course, accidents. Perhaps can the CAPA-tool serve to improve the safety and efficiency of the world fleet by reducing the subjective impact of performance assessment in full-scale simulator training.

7 Conclusions

7.1 Findings

- The CAPA-tool reduced the subjectivity in assessment relative to conventual methods (Ernstsen & Nazir, in review).
- The development of assessment tools in maritime shipping and navigation needs dedicated attention concerning reliability and validity (Ernstsen & Nazir, in review).
- BN and AHP is a viable combination of tools in the development of reliable, valid and flexible assessment tool (Ernstsen, Musharraf, Mallam, et al., 2018; Ernstsen, Musharraf, & Nazir, 2018).
- Pilotage operations must be assessed using multiple dimensions, such as teamwork and technical (Ernstsen, Musharraf, Mallam, et al., 2018; Ernstsen & Nazir, in second review, 2018b; Ernstsen et al., 2016).
- Teams in pilotage operations differ slightly from the rest of the voyage: a critical member is unfamiliar to the rest. Teamwork factors such as adaptability and different types of shared mental models require a sustained working relationship to develop, which could explain why these concept were less mentioned in the study (Ernstsen & Nazir, in second review).
- Further studies are required to investigate the reliability of assessing teamwork (Ernstsen & Nazir, in review).
- The experiment study failed to find evidence of construct validity (Ernstsen & Nazir, in review).

7.2 Stakeholder recommendations

- Formative and summative assessment is of principal importance for education facilities as it will contribute to strengthen the student's learning and the design of future training courses.
- Evidence from Article 4 suggests that the conventional assessment methods are not reliable. The CAPA-tool in its current state could help to structure the performance assessment of technical navigation competencies, but results must be interpreted with caution.
- Training data is becoming increasingly valuable. The structured assessment framework enables the evaluators to develop a database that can be used to further refine and optimise the instrument's precision.
- This could be considered a step towards automated assessment, which is argued to be a needed part in the future of seafaring.
- Automated assessment can eventually, in a longer time-perspective, enable the use of real-time navigation assessment.

References

Accident Investigation Board, N. (2010a). Crete Cement—IMO NO. 9037161, Grounding at Aspond Island in the Oslo Fjord, Norway, on 19 November 2008. *Report Sjø*, *1*.

Accident Investigation Board, N. (2010b). Report on Marine Accident Federal Kivalina-IMO NO. 9205885 Grounding at Årsundøya, Norway 6 October 2008. *Report Sjø*, *1*.

Accident Investigation Board, N. (2012). Report on Investigation Into Marine Accident M/V Godafoss V2PM7 Grounding in Løperen, Hvaler on 17 February 2011. *Report Sjø*, *1*.

Accident Investigation Board, N. (2018). Preliminary marine accident report – collision between the frigate 'KNM Helge Ingstad' and the oil tanker 'Sola TS' on 8 november 2018, outside the Sture terminal in Hjeltefjorden in Hordaland county. *Report Sjø*, *1*.

Ackerman, P. L. (2014). Nonsense, common sense, and science of expert performance: Talent and individual differences. *Intelligence*, *45*, 6–17.

Alexander, D. (2002). Making the Case for Ergonomics. *Proceedings of the Association of Canadian Ergonomics. Alberta, Canada*.

Ali, A. (2006). Simulator instructor-STCW requirements and reality. *Pomorstvo: Scientific Journal of Maritime Research*, *20*(2), 23–32.

Ancona, D. G., & Caldwell, D. F. (1992). Demography and design: Predictors of new product team performance. *Organization Science*, *3*(3), 321–341.

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355.

Andresen, M., Domsch, M. E., & Cascorbi, A. H. (2007). Working unusual hours and its relationship to job satisfaction: A study of European maritime pilots. *Journal of Labor Research*, *28*(4), 714–734.

Armstrong, R., Hall, B. J., Doyle, J., & Waters, E. (2011). 'Scoping the scope'of a cochrane review. *Journal of Public Health*, *33*(1), 147–150.

Ashmawy, E. (2009). Effective Implementation of Safety Management System (SMS): An Overview of the Role of the Human Element. MET Trends in the XXI Century: Shipping Industry and Training Institutions in the global environment–area of mutual interests and cooperation. *Proceedings of the 2009 IAMU General Assembly in St. Petersburg, St. Petersburg, Russia: Admiral Makarov State Maritime Academy. S*, 246–255.

Ashmawy, E. (2012). The maritime industry and the human element phenomenon. *Proc. The 13th Annual General Assembly of the IAMU*.

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine*, *26*(4), 217–238.

Baldauf, M., Dalaklis, D., & Kataria, A. (2016). *Team training in safety and security via simulation: A practical dimension of maritime education and training*.

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology*, *41*(1), 63–105.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, *50*(2), 248–287. https://doi.org/10.1016/0749-5978(91)90022-l

Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, *46*(3), 610.

Bell, B. S., Tannenbaum, S. I., Ford, J. K., Noe, R. A., & Kraiger, K. (2017). 100 years of training and development research: What we know and where we should go. *Journal of Applied Psychology*, *102*(3), 305.

Bell, S. T., Brown, S. G., Colaneri, A., & Outland, N. (2018). Team composition and the ABCs of teamwork. *American Psychologist*, *73*(4), 349.

Bhattacharya, Y. (2015). Employee engagement in the shipping industry: A study of engagement among Indian officers. *WMU Journal of Maritime Affairs*, *14*(2), 267–292. https://doi.org/10.1007/s13437-014-0065-x

BIMCO. (2015). *The global supply and demand for seafarers in 2015*. Retrieved from http://www.ics-shipping.org/docs/default-source/resources/safety-security-and-operations/manpower-report-2015-executive-summary.pdf?sfvrsn=16

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management*, *36*(4), 1065–1105.

Bowen, D. E., Ledford Jr, G. E., & Nathan, B. R. (1991). Hiring for the organization, not the job. *Academy of Management Perspectives*, *5*(4), 35–51.

Boyle, A., & O'Hare, D. (2003). *Finding appropriate methods to assure quality computer-based assessment development in UK higher education*.

Bruno, K., & Lützhöft, M. (2009). Shore-based pilotage: Pilot or autopilot? Piloting as a control problem. *The Journal of Navigation*, *62*(3), 427–437.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Univ of California Press.

Campbell, D. T. (1963). Experimental and quasi-experimental designs for research on teaching. *Handbook of Research on Teaching*, *5*, 171–246.

Campion, M. A., Medsker, G. J., & Higgs, A. C. (1993). Relations between work group characteristics and effectiveness: Implications for designing effective work groups. *Personnel Psychology*, *46*(4), 823–847. https://doi.org/10.1111/j.1744-6570.1993.tb01571.x

Cannon-Bowers, J., Tannenbaum, S., Salas, E., & Volpe, C. (1995). Defining team competencies: Implications for training requirements and strategies. *Team Effectiveness and Decision Making in Organizations*, 333–380.

Carayon, P., & Buckle, P. (2010). Editorial for special issue of applied ergonomics on patient safety. *Applied Ergonomics*, *5*(41), 643–644.

Chambers, T. P., & Main, L. C. (2015). Symptoms of fatigue and coping strategies in maritime pilotage. *International Maritime Health*, *66*(1), 43–48.

Clarke, C., & Milne, R. (2001). *A national evaluation of the PEACE Investigative Interviewing Course*. Home office London.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Conole, G., & Warburton, B. (2005). A review of computer-assisted assessment. *ALT-J*, *13*(1), 17–31.

Converse, S., Cannon-Bowers, J. A., & Salas, E. (1991). *Team member shared mental models: A theory and some methodological issues*. *35*, 1417–1421. SAGE Publications Sage CA: Los Angeles, CA.

Converse, S., Cannon-Bowers, J. A., & Salas, E. (1993). Shared mental models in expert team decision making. *Individual and Group Decision Making: Current Issues*, *221*.

Cook, D. A., & Hatala, R. (2016). Validation of educational assessments: A primer for simulation and beyond. *Advances in Simulation*, *1*(1), 31.

Dahlstrom, N., Dekker, S., Van Winsen, R., & Nyce, J. (2009). Fidelity and validity of simulator training. *Theoretical Issues in Ergonomics Science*, *10*(4), 305–314.

Darbra, R. M., Crawford, J., Haley, C., & Morrison, R. (2007). Safety culture and hazard risk perception of Australian and New Zealand maritime pilots. *Marine Policy*, *31*(6), 736–745. https://doi.org/10.1016/j.marpol.2007.02.004

de Carvalho, P. V. R., Gomes, J. O., Huber, G. J., & Vidal, M. C. (2009). Normal people working in normal organizations with normal equipment: System safety and cognition in a mid-air collision. *Applied Ergonomics*, *40*(3), 325–340.

De Villiers, J. (2000). *Language and theory of mind: What are the developmental relationships?*

De Winter, J. C. (2013). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation*, *18*(10).

De Winter, J. C., Dodou, D., & Mulder, M. (2012). Training effectiveness of whole body flight simulator motion: A comprehensive meta-analysis. *The International Journal of Aviation Psychology*, *22*(2), 164–183.

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science*, *323*(5910), 66–69.

Devine, D. J. (2002). A review and integration of classification systems relevant to teams in organizations. *Group Dynamics: Theory, Research, and Practice*, *6*(4), 291.

Dyer, J. L. (1984). Team research and team training: A state-of-the-art review. *Human Factors Review*, *26*, 285–323.

Eby, L. T., & Dobbins, G. H. (1997). Collectivistic orientation in teams: An individual and group-level analysis. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *18*(3), 275–295.

Emad, G., & Roth, W. M. (2008). Contradictions in the practices of training for and assessment of competency: A case study from the maritime domain. *Education+ Training*, *50*(3), 260–272.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32–64.

Endsley, M. R. (2017). From Here to Autonomy. *Human Factors*, *59*(1), 5–27. https://doi.org/10.1177/0018720816681350

Entin, E. E., & Serfaty, D. (1999). Adaptive team coordination. *Human Factors*, *41*(2), 312–325. https://doi.org/10.1518/001872099779591196

Ernstsen, J., Mallam, S., & Nazir, S. (to be published). Incidental Memory Recall in Virtual Reality: An Empirical Investigation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

Ernstsen, J., Musharraf, M., Mallam, S., Nazir, S., & Veitch, B. (2018). *Bayesian Network for Assessing Performance in Complex Navigation-A Conceptual Model*. *62*, 1751–1755. SAGE Publications Sage CA: Los Angeles, CA.

Ernstsen, J., & Nazir, S. (in second review). Exploring teamwork in maritime pilotage operations. *Ergonomics*.

Ernstsen, J., & Nazir, S. (in review). Performance assessment in full-scale simulators—A case of maritime pilotage operations. *Safety Science*.

Ernstsen, J., & Nazir, S. (2018a). Consistency in the development of performance assessment methods in the maritime domain. *WMU Journal of Maritime Affairs*, *17*(1), 71–90.

Ernstsen, J., & Nazir, S. (2018b). Human error in pilotage operations. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, *12*.

Ernstsen, J., Nazir, S., & Røed, B. K. (2017). Human Reliability Analysis of a Pilotage Operation. *Safety of Sea Transportation*, 293–300. https://doi.org/10.1201/9781315099088-51

Ernstsen, J., Nazir, S., Røed, B. K., & Manca, D. (2016). Systemising performance indicators in the assessment of complex sociotechnical systems. *Chemical Engineering Transactions*, *53*, 187–192. https://doi.org/10.3303/CET1653032

Ernstsen, Musharraf, & Nazir. (2018). *Bayesian Model of Operator Challenges in Maritime Pilotage*. *62*, 1813–1817. SAGE Publications Sage CA: Los Angeles, CA.

Espevik, R., Johnsen, B. H., & Eid, J. (2011). Communication and performance in co-located and distributed teams: An issue of shared mental models of team members? *Military Psychology*, *23*(6), 616.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, *59*, 255–278.

Fischer, U., McDonnell, L., & Orasanu, J. (2007). Linguistic correlates of team performance: Toward a tool for monitoring team functioning during space missions. *Aviation, Space, and Environmental Medicine*, *78*(5), B86–B95.

Fitts, P. M., & Posner, M. I. (1967). *Human performance.*

Flach, J. M. (2012). Complexity: Learning to muddle through. *Cognition, Technology & Work*, *14*(3), 187–197.

Flin, R., Martin, L., Goeters, K.-M., Hormann, H., Amalberti, R., Valot, C., & Nijhuis, H. (2003). Development of the NOTECHS (non-technical skills) system for assessing pilots' CRM skills. *Human Factors and Aerospace Safety*, *3*, 97–120.

Flin, R., & O'Connor, P. (2017). *Safety at the sharp end: A guide to non-technical skills*. CRC Press.

Ford, J. K., & Weissbein, D. A. (1997). Transfer of training: An updated review and analysis. *Performance Improvement Quarterly*, *10*(2), 22–41.

Freelon, D. (2011). Retrieved 29 August 2019, from Dfreelon.org website: http://dfreelon.org/

Gekara, V. O., Bloor, M., & Sampson, H. (2011). Computer-based assessment in safety-critical industries: The case of shipping. *Journal of Vocational Education & Training*, *63*(1), 87–100. https://doi.org/10.1080/13636820.2010.536850

Gladstein, D. L. (1984). Groups in context: A model of task group effectiveness. *Administrative Science Quarterly*, 499–517.

Goodwin, C. (1994). Professional vision. *American Anthropologist*, *96*(3), 606–633.

Graneheim, U. H., & Lundman, B. (2004). Qualitative content analysis in nursing research: Concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today*, *24*(2), 105–112.

Gruenefeld, U., Stratmann, T., Brueck, Y., Hahn, A., Boll, S., & Heuten, W. (2018). Investigations on Container Ship Berthing from the Pilot's Perspective: Accident Analysis, Ethnographic Study, and Online Survey. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, *12*.

Guzzo, R. A., & Shea, G. P. (1992). *Group performance and intergroup relations in organizations.*

Hall, A., & Fagen, R. (1968). *Definition of System. 1 In Buckley, Walter (Ed.) Modern Systems Research for the Behavioral Sciences*.

Hareide, O. S., & Ostnes, R. (2017). Scan pattern for the maritime navigator. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, *11*.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Håvold, J. I. (2005). Safety-culture in a Norwegian shipping company. *Journal of Safety Research*, *36*(5), 441–458.

Håvold, J. I., Nistad, S., Skiri, A., & Ødegård, A. (2015). The human factor and simulator training for offshore anchor handling operators. *Safety Science*, *75*, 136–145.

Hendrick, H. W. (1996). *The ergonomics of economics is the economics of ergonomics*. *40*, 1–10. SAGE Publications Sage CA: Los Angeles, CA.

Hetherington, C., Flin, R., & Mearns, K. (2006). Safety in shipping: The human element. *Journal of Safety Research*, *37*(4), 401–411.

Hodson, P., Saunders, D., & Stubbs, G. (2002). Computer-assisted assessment: Staff viewpoints on its introduction within a new university. *Innovations in Education and Teaching International*, *39*(2), 145–152.

Hollnagel, E., & Woods, D. D. (2005). *Joint cognitive systems: Foundations of cognitive systems engineering*. CRC Press.

Horberry, T., Grech, M., & Koester, T. (2008). *Human factors in the maritime domain*. CRC press.

Hutchins, E. (1995). *Cognition in the Wild*. MIT press.

IEA. (2019). Definition and Domains of Ergonomics | IEA Website. Retrieved 15 July 2019, from https://www.iea.cc/whats/index.html

IMO. (2011). International Convention on Standards of Training, Certification and Watchkeeping for Seafarers (STCW) 1978, as amended in 1995/2010. *International Maritime Organisation, London, UK*.

Jackson, S. E. (1992). Consequences of group composition for the interpersonal dynamics of strategic issue processing. *Advances in Strategic Management*, *8*(3), 345–382.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, *33*(7), 14–26.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. https://doi.org/10.1037/h0034747

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*(23), 2276–2284.

Klein, G. A., Calderwood, R., & Macgregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics*, *19*(3), 462–472.

Kluge, A. (2014). *The acquisition of knowledge and skills for taskwork and teamwork to control complex technical systems: A cognitive and macroergonomics perspective*. Springer.

Kobayashi, H. (2005). Use of simulators in assessment, learning and teaching of mariners. *WMU Journal of Maritime Affairs*, *4*(1), 57–75.

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.

Krippendorff, K. (2011a). Agreement and Information in the Reliability of Coding. *Communication Methods and Measures*, *5*(2), 93–112. https://doi.org/10.1080/19312458.2011.568376

Krippendorff, K. (2011b). *Computing Krippendorff's alpha-reliability*.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Landsberger, H. A. (1958). *Hawthorne Revisited: Management and the Worker, Its Critics, and Developments in Human Relations in Industry.*

Lane, A., Obando-Rojas, B., & Wu, B. (2002). Crewing the International Merchant Fleet, Lloyd's Register. *Fairplay Ltd*.

Lappalainen, J., Kunnaala, V., & Tapaninen, U. (2014). Present pilotage practices in Finland. *WMU Journal of Maritime Affairs*, *13*(1), 77–99. https://doi.org/10.1007/s13437-013-0055-4

Lee, J. D., & Sanquist, T. F. (2000). Augmenting the operator function model with cognitive operations: Assessing the cognitive demands of technological innovation in ship navigation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *30*(3), 273–285.

Liberia. (1967). Liberia: Report on the Stranding of the 'Torrey Canyon' (pollution of the sea by oil). *International Legal Materials*, *6*(3), 480–487.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604.

MacMillan, J., Entin, E. E., & Serfaty, D. (2004). Communication overhead: The hidden cost of team cognition. *Team Cognition: Process and Performance at the Interand Intra-Individual Level. American Psychological Association, Washington, DC.*

Available at Http://Www. Aptima. Com/Publications/2004_MacMillan_EntinEE_Serfaty. Pdf.

Madill, A., Jordan, A., & Shirley, C. (2000). Objectivity and reliability in qualitative analysis: Realist, contextualist and radical constructionist epistemologies. *British Journal of Psychology*, *91*(1), 1–20.

Mallam, S., Ernstsen, J., & Nazir, S. (to be published). Safety in Shipping: Investigating Safety Climate in Norwegian Maritime Workers. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

Mallam, S., Nazir, S., Renganayagalu, S. K., Ernstsen, J., Veie, S., & Edwinson, A. E. (2019). Design of Experiment Comparing Users of Virtual Reality Head-Mounted Displays and Desktop Computers. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (pp. 240–249). Springer International Publishing.

Manca, D., Nazir, S., Colombo, S., & Kluge, A. (2014). Procedure for automated assessment of industrial operators. *Chemical Engineering Transactions*, *36*, 391–396.

Maran, N. J., & Glavin, R. J. (2003). Low-to high-fidelity simulation–a continuum of medical education? *Medical Education*, *37*, 22–28.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26*(3), 356–376.

Mathieu, J. E., Tannenbaum, S. I., Donsbach, J. S., & Alliger, G. M. (2014). A review and integration of team composition models: Moving toward a dynamic and temporal framework. *Journal of Management*, *40*(1), 130–160.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica: Biochemia Medica*, *22*(3), 276–282.

McIntyre, R. M., & Salas, E. (1995). Measuring and managing for team performance: Emerging principles from complex environments. *Team Effectiveness and Decision Making in Organizations*, 9–45.

McKinney Jr, E. H., Barker, J. R., Smith, D. R., & Davis, K. J. (2004). The role of communication values in swift starting action teams: IT insights from flight crew experience. *Information & Management*, *41*(8), 1043–1056.

Messick, S. (1987). Validity. *ETS Research Report Series*, *1987*(2), i–208.

Millán, E., DescalçO, L., Castillo, G., Oliveira, P., & Diogo, S. (2013). Using Bayesian networks to improve knowledge assessment. *Computers & Education*, *60*(1), 436–447.

Mohammed, S., & Dumville, B. C. (2001). Team mental models in a team knowledge framework: Expanding theory and measurement across disciplinary boundaries. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *22*(2), 89–106.

Moorthy, K., Munz, Y., Sarker, S. K., & Darzi, A. (2003). Objective assessment of technical skills in surgery. *Bmj*, *327*(7422), 1032–1037.

Morrison, B., & Morrison, N. M. (2018). *Capturing cognition using the Critical Decision Method*. Presented at the Human Factors and Ergonomics Conference (Human Factors and Ergonomics Society of Australia).

Murdock Jr, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482.

Musharraf, M., Hassan, J., Khan, F., Veitch, B., MacKinnon, S., & Imtiaz, S. (2013). Human reliability assessment during offshore emergency conditions. *Safety Science*, *59*, 19–27. https://doi.org/10.1016/j.ssci.2013.04.001

Musharraf, M., Smith, J., Khan, F., Veitch, B., & MacKinnon, S. (2016). Assessing offshore emergency evacuation behavior in a virtual environment using a Bayesian network approach. *Reliability Engineering & System Safety*, *152*, 28–37.

Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and Organization*, *17*(1), 2–26. https://doi.org/10.1016/j.infoandorg.2006.11.001

Nazir, S., Øvergård, K. I., & Yang, Z. (2015). Towards Effective Training for Process and Maritime Industries. *Procedia Manufacturing*, *3*, 1519–1526. https://doi.org/10.1016/j.promfg.2015.07.409

Nazir, S., Sorensen, L., Øvergård, K. I., & Manca, D. (2015). Impact of training methods on Distributed Situation Awareness of industrial operators. *Safety Science*, *73*, 136–145. https://doi.org/10.1016/j.ssci.2014.11.015

Neisser, U. (1976). Cognition and reality. Principles and implication of cognitive psychology. *San Francisko: WH Freeman and Company*.

Neuendorf, K. A. (2016). *The content analysis guidebook*. Sage.

Neyman, J., & Pearson, E. S. (1933). *The testing of statistical hypotheses in relation to probabilities a priori*. *29*, 492–510. Cambridge University Press.

Nightingale, D., & Cromby, J. (1999). *Social constructionist psychology: A critical analysis of theory and practice*. McGraw-Hill Education (UK).

Nisbett, R. E., & Wilson, T. D. (1977a). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231.

Nisbett, R. E., & Wilson, T. D. (1977b). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*(4), 250.

Norros, L. (2004). Acting under uncertainty. *The Core-Task Analysis in Ecological Study of Work. Espoo, VTT, Finland*.

Norros, L., & Hukki, K. (2003). Utilization of information technology in navigational decision-making. *Cooperative Process Management: Cognition And Information Technology: Cognition And Information Technology*, 77.

O’Connor, P. (2011). Assessing the effectiveness of bridge resource management training. *The International Journal of Aviation Psychology*, *21*(4), 357–374.

Paine, L. (2014). *The sea and civilization: A maritime history of the world*. Atlantic Books Ltd.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Elsevier.

Pickering, C., & Byrne, J. (2014). The benefits of publishing systematic quantitative literature reviews for PhD candidates and other early-career researchers. *Higher Education Research & Development*, *33*(3), 534–548. https://doi.org/10.1080/07294360.2013.841651

Podgórski, D. (2015). Measuring operational performance of OSH management system–A demonstration of AHP-based selection of leading key performance indicators. *Safety Science*, *73*, 146–166.

Pohl, R., & Pohl, R. F. (2004). *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press.

Porathe, T. (2019). *Interaction Between Manned and Autonomous Ships: Automation Transparency*. Presented at the Proceedings of the 1st International Conference on Maritime Autonomous Surface Ships.

Potter, J., & Hepburn, A. (2005). Qualitative interviews in psychology: Problems and possibilities. *Qualitative Research in Psychology*, *2*(4), 281–307.

Praetorius, G., & Hollnagel, E. (2014). Control and resilience within the maritime traffic management domain. *Journal of Cognitive Engineering and Decision Making*, *8*(4), 303–317.

Priest, H. A., Burke, C. S., Munim, D., & Salas, E. (2002). *Understanding team adaptability: Initial theoretical and practical considerations*. *46*, 561–565. SAGE Publications Sage CA: Los Angeles, CA.

Progoulaki, M., & Roe, M. (2011). Dealing with multicultural human resources in a socially responsible manner: A focus on the maritime industry. *WMU Journal of Maritime Affairs*, *10*(1), 7–23.

Rafferty, L. A., Stanton, N. A., & Walker, G. H. (2010). The famous five factors in teamwork: A case study of fratricide. *Ergonomics*, *53*(10), 1187–1204.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, *28*(1), 4–13.

Ramsden, P. (1997). The context of learning in academic departments. *The Experience of Learning*, *2*, 198–216.

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*, (3), 257–266.

Renganayagalu, S. kumar, Mallam, S., Nazir, S., Ernstsen, J., & Haavardtun, P. (to be published). Impact of simulation fidelity on student self-efficacy and perceived skill development in engine room simulators. *Safety of Sea Transportation: Proceedings of the 12th International Conference on Marine Navigation and Safety of Sea Transportation*.

Roberts, C., Newble, D., Jolly, B., Reed, M., & Hampton, K. (2006). Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Medical Teacher*, *28*(6), 535–543.

Ropohl, G. (1999). Philosophy of socio-technical systems. *Techné: Research in Philosophy and Technology*, *4*(3), 186–194.

Rowley, J. (2012). Conducting research interviews. *Management Research Review*, *35*(3/4), 260–271.

Ruggeberg, B. (2007). *A consultant's perspective on doing competencies well: Methods, models, and lessons*. Presented at the Doing competencies well, Symposium presented at the Annual Conference of the Society for Industrial and Organizational Psychology, New York.

Rumawas, V. (2016). *Human factors in ship design and operation: Experiential learning*.

Saaty, R. W. (1987). The analytic hierarchy process—What it is and how it is used. *Mathematical Modelling*, *9*(3–5), 161–176.

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, *15*(3), 234–281.

Saaty, T. L. (1988). What is the analytic hierarchy process? In *Mathematical models for decision support* (pp. 109–121). Springer.

Saaty, T. L. (1990). *Multicriteria decision making: The analytic hierarchy process: Planning, priority setting resource allocation*.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144.

Safrit, M. J., & Wood, T. M. (1989). *Measurement concepts in physical education and exercise science*. Human Kinetics Books Champaign.

Salas, E. (1992). Toward an understanding of team performance and training. *Teams: Their Training and Performance*, 3–29.

Salas, E., Burke, C. S., Fowlkes, J. E., & Priest, H. A. (2004). On measuring teamwork skills. *Comprehensive Handbook of Psychological Assessment*, *4*, 427–442.

Salas, E., Sims, D., & Burke, S. (2005). Is there a "big five" in teamwork? *Small Group Research*, *36*(5), 555–599. https://doi.org/10.1177/1046496405277134

Salas, E., Sims, D. E., & Klein, C. (2004). Cooperation at work. *Encyclopedia of Applied Psychology*, *1*, 497–505.

Salas, E., Wilson, K. A., Burke, C. S., & Wightman, D. C. (2006). Does crew resource management training work? An update, an extension, and some critical needs. *Human Factors*, *48*(2), 392–412.

Sampson, H. (2004). Romantic rhetoric, revisionist reality: The effectiveness of regulation in maritime education and training. *Journal of Vocational Education and Training*, *56*(2), 245–267.

Scriven, M. (1967). The methodology of evaluation. In *Perspectives on curriculum evaluation*. Chicago: Rand McNally and Co.

Sellberg, C. (2017). Simulators in bridge operations training and assessment: A systematic review and qualitative synthesis. *WMU Journal of Maritime Affairs*, *16*(2), 247–263.

Sellberg, C., & Lundin, M. (2017). Demonstrating professional intersubjectivity: The instructor's work in simulator-based learning environments. *Learning, Culture and Social Interaction*, *13*, 60–74.

Sellberg, C., & Susi, T. (2014). Technostress in the office: A distributed cognition perspective on human–technology interaction. *Cognition, Technology & Work*, *16*(2), 187–201.

Shah, J., & Breazeal, C. (2010). An empirical analysis of team coordination behaviors and action planning with application to human–robot teaming. *Human Factors*, *52*(2), 234–245.

Sharma, A., Nazir, S., & Ernstsen, J. (2019). Situation awareness information requirements for maritime navigation: A goal directed task analysis. *Safety Science*, *120*, 745–752.

Siegel, A., & Federman, P. (1973). Communications content training as an ingredient in effective team performance. *Ergonomics*, *16*(4), 403–416.

Simsarian Webber, S. (2002). Leadership and trust facilitating cross-functional team success. *Journal of Management Development*, *21*(3), 201–214. https://doi.org/10.1108/02621710210420273

Singer, T., & Fehr, E. (2005). The neuroeconomics of mind reading and empathy. *American Economic Review*, *95*(2), 340–345.

Singh, A. S., & Masuku, M. B. (2014). Sampling techniques & determination of sample size in applied statistics research: An overview. *International Journal of Economics, Commerce and Management*, *2*(11), 1–22.

Sleire, H., & Dale, E. (2009). *The Shipping KPI Standard*. *2009*, 6–7.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, *6*(6), 649–744.

Squire, L. R. (1992). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of Cognitive Neuroscience*, *4*(3), 232–243.

Stanovich, K. E., & Toplak, M. E. (2012). Defining features versus incidental correlates of Type 1 and Type 2 processing. *Mind & Society*, *11*(1), 3–13. https://doi.org/10.1007/s11299-011-0093-6

Stanton, N. A., & Bessell, K. (2014). How a submarine returns to periscope depth: Analysing complex socio-technical systems using Cognitive Work Analysis. *Applied Ergonomics*, *45*(1), 110–125.

Stanton, N. A., & Salmon, P. M. (2011). Planes, trains and automobiles: Contemporary ergonomics research in transportation safety. *Applied Ergonomics*, *42*(4), 529–532.

Stopford, M. (2009). *Maritime economics 3e*. Routledge.

Sundstrom, E. (1999). The challenges of supporting work team effectiveness. *Supporting Work Team Effectiveness*, *3*, 23.

Swezey, R. W., & Salas, E. (1992). *Guidelines for use in team-training development.*

Taras, M. (2005). Assessment–summative and formative–some theoretical reflections. *British Journal of Educational Studies*, *53*(4), 466–478.

Taylor, D. H. (1998). Rules and regulations in maritime collision avoidance: New directions for bridge team training. *The Journal of Navigation*, *51*(1), 67–72.

Taylor, P. J., Driscoll, M. P. O., & Binning, J. F. (1998). A new integrated framework for training needs analysis. *Human Resource Management Journal*, *8*(2), 29–50.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

UNCTAD. (2018). *Review of Maritime Transport* (UNCTAD/RMT/2018 No. 978-92-1-112928–1). United Nations.

van Westrenen, F., & Praetorius, G. (2014). Maritime traffic management: A need for central coordination? *Cognition, Technology & Work*, *16*(1), 59–70.

Vederhus, L., Ødegård, A., Nistad, S., & Håvold, J. I. (2018). Perceptions of demanding work in maritime operations. *Safety Science*, *110*, 72–82.

Vicente, K. J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. CRC Press.

Wang, H., Sheng, X., Wang, S., Chen, L., Yuan, Z., & Wu, Q. (2017). Numerical study on water depth effects on hydrodynamic forces acting on berthing ships. *Journal of Shanghai Jiaotong University (Science)*, *22*(2), 198–205.

Whittemore, R., Chao, A., Jang, M., Minges, K. E., & Park, C. (2014). Methods for knowledge synthesis: An overview. *Heart & Lung*, *43*(5), 453–461.

Wickens, C. D., Gordon, S. E., Liu, Y., & Lee, J. (1998). *An introduction to human factors engineering*.

Wickens, C. D., Stokes, A., Barnett, B., & Hyman, F. (1993). The effects of stress on pilot judgment in a MIDIS simulator. In *Time pressure and stress in human judgment and decision making* (pp. 271–292). Springer.

Wilcox, T. (2000). STCW-95: Officer in charge of a navigational watch. *Marine Safety Council Proceedings*, *57*(1), 39–41.

Wild, C. R. J. (2011). The paradigm and the paradox of perfect pilotage. *The Journal of Navigation*, *64*(1), 183–191.

Willig, C. (2013). *Introducing qualitative research in psychology*. McGraw-hill education (UK).

Wilson, K. A., Salas, E., Priest, H. A., & Andrews, D. (2007). Errors in the heat of battle: Taking a closer look at shared cognition breakdowns through teamwork. *Human Factors*, *49*(2), 243–256.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*(1), 117.

Woods, D. D., Johannesen, L. J., Cook, R. I., & Sarter, N. B. (1994). *Behind human error: Cognitive systems, computers and hindsight*. DTIC Document.

Yudkowsky, E. (2008). Cognitive biases potentially affecting judgment of global risks. *Global Catastrophic Risks*, *1*(86), 13.

Yule, S, Flin, R., Paterson-Brown, S., & Maran, N. (2006). Non-technical skills for surgeons in the operating room: A review of the literature. *Surgery*, *139*(2), 140–149.

Yule, Steven, Rowley, D., Flin, R., Maran, N., Youngson, G., Duncan, J., & Paterson-Brown, S. (2009). Experience matters: Comparing novice and expert ratings of non-technical skills using the NOTSS system. *ANZ Journal of Surgery*, *79*(3), 154–160.

Zakrzewski, S., & Steven, C. (2000). A model for computer-based assessment: The Catherine wheel principle. *Assessment & Evaluation in Higher Education*, *25*(2), 201–215.

Zoogah, D. B., Noe, R. A., & Shenkar, O. (2015). Shared mental model, team communication and collective self-efficacy: An investigation of strategic alliance team effectiveness. *International Journal of Strategic Business Alliances*, *4*(4), 244–270.

Doctoral dissertation no. 64
2020

—

Reducing the subjective impact in maritime simulator assessment

A performance assessment tool for maritime pilotage operations

—

Jørgen Ernsten

—

ISBN: 978-82-7860-423-6 (print)
ISBN: 82-7860-425-8 (online)

usn.no