TRANSNAV
http://www.transnav.eu
the International Journal
on Marine Navigation
and Safety of Sea Transportation
Volume 12
Number 3
September 2018

DOI: 10.12716/1001.12.03.20

A proposed Evidential Reasoning (ER) Methodology for Quantitative Assessment of Non-Technical Skills (NTS) Amongst Merchant Navy Deck Officers in a Ship's Bridge Simulator Environment

F. Saeed
Higher Colleges of Technology, Abu Zabi, United Arab Emirates

A. Bury, S. Bonsall & R. Riahi
Liverpool John Moores University, Liverpool, United Kingdom

ABSTRACT: Ship's bridge simulators are very popular in the worldwide training and assessment of merchant navy deck officers. The examiners of simulator courses presently do not have a method to quantitatively assess the performance of a group or an individual. Some examiners use checklists and others use their gut feeling to grade competence. In this paper a novel methodology is established that uses the Evidential Reasoning algorithm to quantitatively assess the Non-Technical Skills (NTS) of merchant navy officers. To begin with, interviews were conducted with experienced deck officers to develop the taxonomy and behavioural markers that would be used in the assessment process. A random selection of students studying towards their Chief Officer's Certificate of Competency were recruited to have their NTS to be observed in a ship's bridge simulator. The participant's behaviour was rated against five criteria and the subsequent data was entered into the Evidential Reasoning algorithm to produce a crisp number. The results that were generated demonstrate that this approach provides a reliable method to quantitatively assess the NTS performance of merchant navy officers in a simulated bridge environment.

1 INTRODUCTION

NTS are those specific human competencies such as leadership, teamwork, situation awareness and decision making, which affect the likelihood of human error occurring and the severity of its impact (Flin et al., 2003). The four main NTS are subdivided into two categories; social and cognitive. Social skills are those which are easily observable *i.e* leadership and team-working. Cognitive skills are those which are difficult to observe *i.e* situation awareness and decision making (Flin et al., 2003).

Simulator training has proven to be very successful in the training of personnel for operating in high risk domains (Kozuba and Bondaruk, 2014; Wanger et al., 2013; Balci et al., 2014). Many safety critical industries, such as aviation and anaesthesia, have now adapted simulation as the recommended method of NTS training and its effectiveness has been tested in various pieces of research across the globe worldwide (Winter et al., 2012; Michael et al., 2014).

The technology has also been adopted for training and assessments in the maritime sector. The mathematical model of a ship created on a computer graphically displays the ship and its movement through the water nearly in a realistic manner and helps learners to learn effectively (Mohovic et al., 2012). The training provided through this medium has many benefits such as the ability to navigate vessels through restricted waters, deal with emergency or crisis situations or use various navigational aids (Pelletier, 2006). The biggest advantage of providing training by simulator is the ability to create various scenarios in different

meteorological conditions in different sea areas using different target ships (Sniegocki, 2005).

Simulator training is now being used as a compulsory training element of the Officer of the Watch (OOW) and Chief Mate's course. At the OOW level the course is called NAEST (O) (Navigation Aids and Equipment Simulator Training – Operational) and at chief mate's level NAEST (M) (Navigation Aids and Equipment Simulator Training – Management). The NAEST (O) course is a basic level course where the use of equipment, basic watch keeping and navigation skills are taught to students undertaking the OOW course. Whereas NAEST (M) is a management level course where advanced navigation skills are taught (Wall, 2015).

Presently simulator assessors do not have any method to quantitatively assess the NTS competence of deck officers. They normally use their gut feeling to gauge the competence of a candidate.

2 METHODS

The aim of this research is to develop a methodology for quantitatively assessing the NTS of merchant navy deck officers in a ship's bridge simulator. To achieve this, the following steps were undertaken :

1 Develop a taxonomy for deck officers' NTS. To assign a weight to each different criterion, questionnaires were designed to assign the possible values for ranking each different criterion through meetings and interviews with the experienced deck officers. The ranks/weights assigned by experts were aggregated by the AHP method.
2 Develop a behavioural markers' assessment framework based on the taxonomy of deck officers' NTS.
3 Simulator scenario developed and volunteer chief officer students recruited.
4 Simulator observations conducted with volunteer students and each BM was awarded a weight by assessor.
5 ER Algorithm and UV method used to calculate the final crisp number of the performance.

3 DEVELOP A TAXONOMY FOR DECK OFFICERS NTS (STEP 1)

To develop a taxonomy of deck officers NTS, a series of interviews were conducted with experienced deck officers at management level to help identify the key skills to be included. A semi-structured method of interviewing was carried out to extract maximum information from the interviewee. The aim of each interview was to identify the non-technical aspect of a deck officer's role in a crisis situation on the bridge of a ship and the skills needed for this, e.g. thinking and team working skills, decision making, situation awareness and leadership.

The interview was divided into three parts:

– Part 1: Performance example – The interviewee was asked to describe a real case from his career that was particularly challenging which tested his NTS. The example could be a real critical incident/near miss or a normal case where experience and NTS were a significant outcome. The interviewee was asked in advance if he could think of this example before the interview. This case was then discussed to identify the most significant NTS components.
– Part 2: Distinguishing skills – The interviewee was asked to think about the skills which are necessary for the effective performance of a deck officer involved in a crisis situation on the bridge of a ship.
– Part 3: Weighting task – The interviewee was asked to assign a weight to each of the NTS taxonomy elements.

Approximate times for the three interview parts were: Part 1 – 45 minutes, Part 2 – 15 minutes, Part 3 – 15 minutes. All the given information was held in confidence and is kept as anonymous.

3.1 *Pilot Interview*

To support the development of the interview schedule, a pilot interview was undertaken with a senior deck officer. This took place at an early stage to help make minor changes to the interview questionnaire. This questionnaire was adapted from the study of 'Identification and measurement of anaesthetists' NTS (Fletcher et al. 2003b). The pilot interview was recorded and subsequently utilised by the research team to ensure that the necessary information was being obtained from the interviews.

3.2 *Identifying Participants*

The first criterion for the selection of the participants was that they must hold a Master Mariner Certificate of Competency. The other criterion for taking part in the study was that the interviewees volunteered to take part. Fletcher et al. (2003b) argues that those people who are very interested in human factors will be more inclined to volunteer and this might lead to potential biases. However, given the sensitivity of the information being discussed, it would be unethical to interview unwilling participants. The researcher in this project visited the World Maritime University, Malmo, to conduct interviews with experienced master mariners pursuing further studies. The researcher's aim was to conduct 10-15 interviews for this research but could only manage 12 interviews in total.

3.3 *Data Analysis*

Based on a review of the existing literature and with the help of the information collected from experienced seafarers through the interview process, a generic decision making model was generated (Figure 1), the data gathered during the interviews, was carefully reviewed and a weight assigned to each criterion using the mathematical decision making method known as the Analytical Hierarchy Process (AHP). The process of evaluating weight of a criterion is presented in the following subsection.

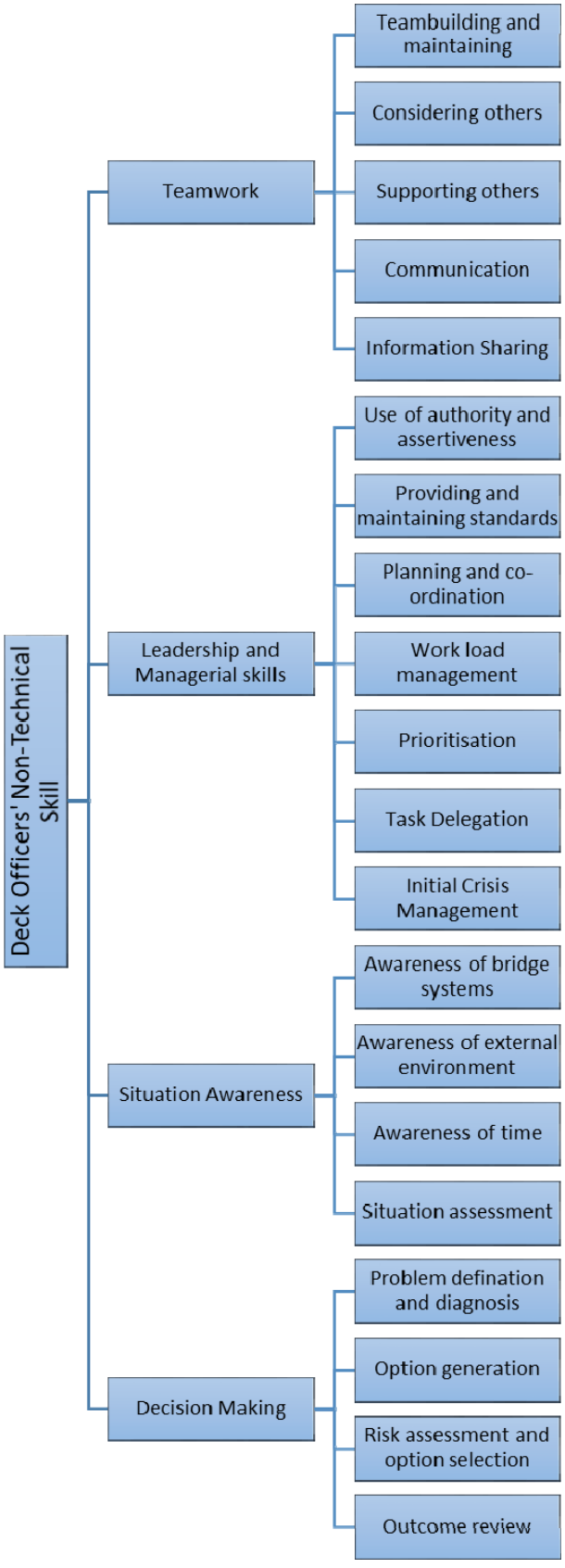

Figure 1. Deck Officers' Non-technical Skills Taxonomy

3.3.1 *The AHP method*

The AHP was pioneered by Saaty and is often referred to as the Saaty method (Coyle, 2004). The method is popular and widely used in decision making and rating tasks. It is a multi-criteria decision making (MCDM) method that helps the decision-maker to make the right decision in a complex situation (Ishizaka and Labib, 2009). AHP case applications range from choice of career through to planning a port development (Coyle, 2004).

Riahi et al. (2012) has used Saaty's quantified judgements on pairs of attributes A_i and A_j represented by an n-by-n matrix D. The entries a_{ij} are defined by the following entry rules.

Rule 1. If $a_{ij} = \alpha$, then $a_{ji} = 1/\alpha$, $\alpha \neq 0$

Rule 2. If A_i is judged to be of equal relative importance as A_j, then $a_{ij} = a_{ji} = 1$

$$D = \begin{bmatrix} 1 & a_{12} & \dots & a_{1n} \\ 1/a_{12} & 1 & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ 1/a_{1n} & 1/a_{2n} & \dots & 1 \end{bmatrix}$$

where i, j = 1, 2, 3, …, n and each a_{ij} is relative importance of attribute A_i to attribute A_j.

Having recorded the quantified judgments of comparison on pair (A_i, A_j) as the numerical entry a_{ij} in the matrix D, what is left is to assign to the n contingencies A_1, A_2, …, A_n a set of numerical weights w_1, w_2, …, w_n that should reflect the recorded judgements. Generally weights w_1, w_2,…, w_n can be calculated by using the following equation;

$$\omega_k = \frac{1}{n}\sum_{j=1}^{n}\frac{a_{kj}}{\sum_{i=1}^{n} a_{ij}}\left(k = 1,2,3,\dots,n\right) \quad (1)$$

where a_{ij} represents the entry of row i and column j in a comparison matrix of order n.

The weight vector of the comparison matrix will provide the priority order but it cannot confirm the consistency of the pairwise judgement. The AHP provides a measure of the consistency of the pairwise comparisons by computing a Consistency Ratio (CR) (Riahi et al., 2012). The CR is devised in such a way that a value less than 0.10 is deemed consistent in that a decision maker should review the pairwise judgements if the resultant value is more than 0.10.

The CR value is calculated according to the following equations:

$$CR = \frac{CI}{RI} \quad (2)$$

$$CI = \frac{\lambda_{max} - n}{n - 1} \quad (3)$$

$$\lambda_{max} = \frac{\sum_{j=1}^{n}[(\sum_{k=1}^{n} w_k a_{jk}) / w_j]}{n} \quad (4)$$

where CI is the Consistency Index, RI is the average random index (Table 4.7), *n* is the matrix

order and λ_{max} is the maximum weight value of the *n-by-n* comparison matrix D.

The following numerical example shows the method of evaluation of weights of main criteria (i.e. Situation Awareness, Decision Making, Leadership and Team Work) by an anonymous expert judgement (Table 2).

Table 1. Value of *RI* versus matrix order (Saaty, 1990)

n	RI
1	0
2	0
3	0.58
4	0.9
5	1.12
6	1.24
7	1.32
8	1.41
9	1.45
10	1.49

$$D=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

The matrix for main criterion was obtained from the table 2 as follows:

$$\begin{array}{c} \\ SA \\ DM \\ LS \\ TW \end{array} \begin{array}{c} \begin{array}{cccc} SA & DM & LS & TW \end{array} \\ D=\begin{bmatrix} 1 & 1 & 1/3 & 2 \\ 1 & 1 & 1 & 3 \\ 3 & 1 & 1 & 3 \\ 1/2 & 1/3 & 1/3 & 1 \end{bmatrix} \end{array}$$

Weights of main criteria are calculated using equation 1:

$$\omega_1=\frac{1}{n}\left(\frac{a_{11}}{(a_{11}+a_{21}+a_{31}+a_{41})}+\frac{a_{12}}{(a_{12}+a_{22}+a_{32}+a_{42})}+\frac{a_{13}}{(a_{13}+a_{23}+a_{33}+a_{43})}+\frac{a_{14}}{(a_{14}+a_{24}+a_{34}+a_{44})}\right)$$

$$\omega_1=\frac{1}{4}\left(\frac{1}{(1+1+3+0.5)}+\frac{1}{(1+1+1+0.3333)}+\frac{0.3333}{(0.3333+1+1+0.3333)}+\frac{2}{(2+3+3+1)}\right)$$

$$\omega_1=0.207260$$

$$\omega_2=\frac{1}{n}\left(\frac{a_{21}}{(a_{11}+a_{21}+a_{31}+a_{41})}+\frac{a_{22}}{(a_{12}+a_{22}+a_{32}+a_{42})}+\frac{a_{23}}{(a_{13}+a_{23}+a_{33}+a_{43})}+\frac{a_{24}}{(a_{14}+a_{24}+a_{34}+a_{44})}\right)$$

$$\omega_2=\frac{1}{4}\left(\frac{1}{(1+1+3+0.5)}+\frac{1}{(1+1+1+0.3333)}+\frac{1}{(0.3333+1+1+0.3333)}+\frac{3}{(2+3+3+1)}\right)$$

$$\omega_2=0.297538$$

$$\omega_3=\frac{1}{n}\left(\frac{a_{31}}{(a_{11}+a_{21}+a_{31}+a_{41})}+\frac{a_{32}}{(a_{12}+a_{22}+a_{32}+a_{42})}+\frac{a_{33}}{(a_{13}+a_{23}+a_{33}+a_{43})}+\frac{a_{34}}{(a_{14}+a_{24}+a_{34}+a_{44})}\right)$$

$$\omega_3=\frac{1}{4}\left(\frac{3}{(1+1+3+0.5)}+\frac{1}{(1+1+1+0.3333)}+\frac{1}{(0.3333+1+1+0.3333)}+\frac{3}{(2+3+3+1)}\right)$$

$$\omega_3=0.388447$$

$$\omega_4=\frac{1}{n}\left(\frac{a_{41}}{(a_{11}+a_{21}+a_{31}+a_{41})}+\frac{a_{42}}{(a_{12}+a_{22}+a_{32}+a_{42})}+\frac{a_{43}}{(a_{13}+a_{23}+a_{33}+a_{43})}+\frac{a_{44}}{(a_{14}+a_{24}+a_{34}+a_{44})}\right)$$

$$\omega_4=\frac{1}{4}\left(\frac{0.5}{(1+1+3+0.5)}+\frac{0.3333}{(1+1+1+0.333)}+\frac{0.3333}{(0.3333+1+1+0.3333)}+\frac{1}{(2+3+3+1)}\right)$$

$$\omega_4=0.106755$$

Table 2: Anonymous expert judgements
Goal: To Select the most important non-technical skills for deck Officers
Situation Awareness

How important is .. 'Situation Awareness' compared to	Unimportant								Equally Important	Important							
	1/9	1/8	1/7	1/6	1/5	1/4	1/3	1/2	1	2	3	4	5	6	7	8	9
Decision Making									x								
Leadership							x										
Teamwork										x							

Decision Making

How important is .. 'Decision Making' compared to	Unimportant								Equally Important	Important							
	1/9	1/8	1/7	1/6	1/5	1/4	1/3	1/2	1	2	3	4	5	6	7	8	9
Leadership									x								
Teamwork											x						

Leadership

How important is .. 'Leadership' compared to	Unimportant								Equally Important	Important							
	1/9	1/8	1/7	1/6	1/5	1/4	1/3	1/2	1	2	3	4	5	6	7	8	9
Teamwork											x						

The weight values are found as 0.207260 (ω_1), 0.297538 (ω_2), 0.388447 (ω_3) and 0.106755 (ω_4). Consistency ratio is calculated by using equations 2, 3, 4.

Based on equation 4, λ_{max} was calculated as follows:

ω_{1x} = (1 × 0.207260) + (1 × 0.297538) + (0.333333 × 0.388447) + (2 × 0.106755) = 0.847790

ω_{2x} = (1 × 0.207260) + (1 × 0.297538) + (1 × 0.388447) + (3 × 0.106755) = 1.21351

ω_{3x} = (3 × 0.207260) + (1 × 0.297538) + (1 × 0.388447) + (3 × 0.106755) = 1.62803

ω_{4x} = (0.5 × 0.20726) + (0.33 × 0.297538) + (0.33 × 0.388447) + (1 × 0.106755) = 0.43905

$$\lambda_{max} = \frac{\left(\frac{0.847790}{0.207260}\right)+\left(\frac{1.21351}{0.297538}\right)+\left(\frac{1.62803}{0.388447}\right)+\left(\frac{0.43905}{0.106755}\right)}{4} = 4.118196$$

The mean value for λ_{max} is 4.118196. If any of the λ_{max} turns out to be less than n, which is 4 in this case, then there is an error in the calculation, which requires a thorough check.

The CI is calculated as follows;

$$CI = \frac{\lambda_{max} - n}{n-1} = \frac{4.118196 - 4}{4-1} = 0.03939$$

Based on table 1, the Random Index (RI) for 4 criteria is 0.9. As a result, the CR value was calculated as follows;

$$CR = \frac{CI}{CR} = \frac{0.03939}{0.9} = 0.04376$$

The CR value for the main criteria was found to be 0.04376. A CR value of less than or equal to 0.1 indicates that judgements are acceptable (Saaty, 1980). As a result, the consistency of pair-wise comparisons for the main criteria, are acceptable. The same calculation technique was applied to obtain weights for each sub-criterion and to check the consistency of the expert opinions.

3.3.2 *Geometric Mean Method*

AHP initially was developed as a decision making tool for individual decision makers but by the use of the geometric mean method individual pairwise comparison metrics of any number of experts can be aggregated (Aull-Hyde et al., 2006) as follows:

$$\text{GeometricMean}_{ij} = [e_{1ij}.e_{2ij}.e_{3ij}\ldots e_{kij}]^{\frac{1}{k}} \quad (5)$$

where, e_{kij} is the k^{th} expert judgement on pair of attributes A_i and A_j.

3.3.3 *Knowledge Representation*

Data was collected by conducting interviews with 12 experienced senior deck officers both in UK and Malmo, Sweden. Only eight participants' results were considered for this study as the remaining four participants' weighting data was inconsistent in light of the AHP formula. Figure 2 shows the weights of all elements of the NTS.

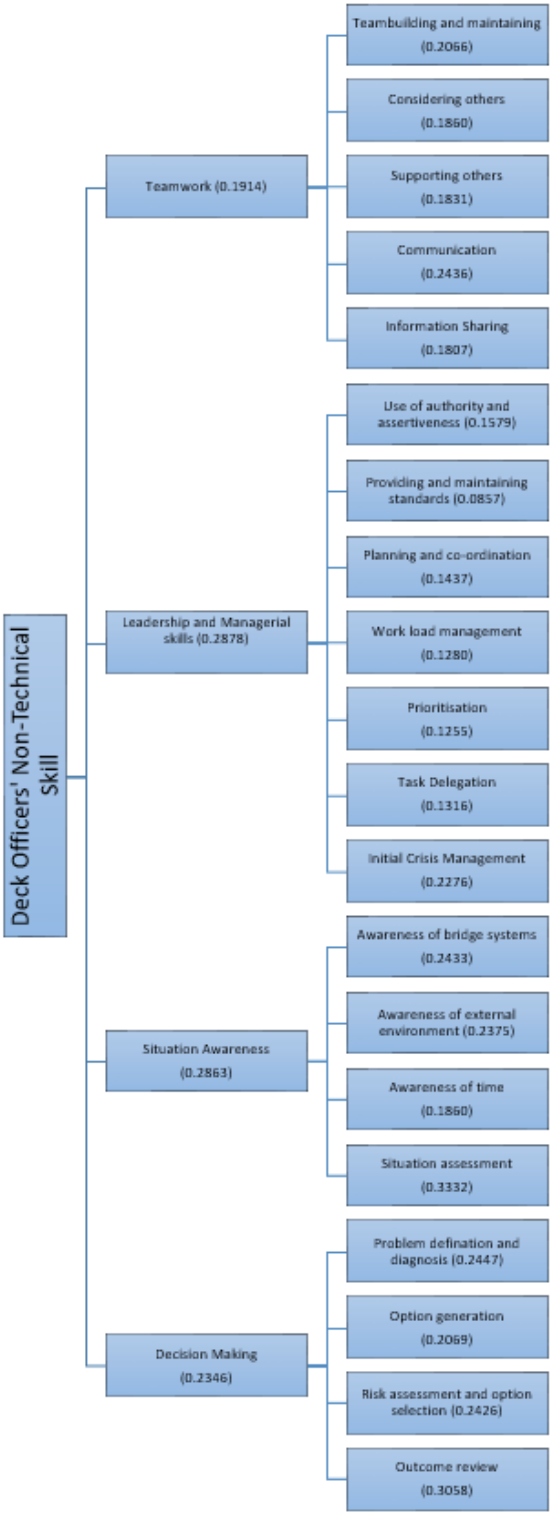

Figure 2. Deck Officers' Non-technical Skills Taxonomy (With resultant weights)

4 DEVELOPMENT OF BEHAVIOURAL MARKERS (STEP 2)

Behavioural marker systems are used for training and assessments of the participants in the simulators and were first developed in the aviation industry (Helmreich et al., 1999). Later on other safety critical industries such as anaesthesia and nuclear power generation have developed their own behavioural marker systems.

Klampfer et al. (2001) proposed the following for designing good behaviour marker systems:
- Validity: in relation to performance outcome.
- Reliability: inter-rater reliability, internal consistency.
- Sensitivity: in relation to levels of performance.
- Transparency: the observer understands the performance criteria against which they are being rated, availability of reliability and validity data.
- Usability: easy to train, simple framework, easy to understand, domain appropriate language, sensitive to rater workload, easy to observe.

Klampfer et al. (2001) further suggest that behavioural marker systems are limited because they "cannot capture every aspect of performance and behaviour" due to the:
- Limited occurrence of some behaviours such as conflict resolution.
- Limitation of human observers such as distraction or overload (e.g. in complex situations, or when observing large teams)

In developing behavioural markers systems for scrub practitioners' NTS (SPLINTS system) Mitchell et al. (2013) established the following design criteria:
- Focus on the skills that are observable from behaviour.
- Be set as a hierarchical structure with three levels of description; category, element, and behaviour.
- Use active verbs for skills and understandable language for definitions.
- Show a simple structure and layout with a rating scale that fits on one page that it can be easily used.

The behavioural marker assessment framework must, as far as possible, be designed to ensure that it is capable of capturing the fullest context of the environment in which the assessment is taking place (Gatfield, 2008). Behavioural markers are a valuable tool in assessing or observing a participant's technical and NTS in the real world or in the simulator.

A review of behaviour marker systems in use in other safety critical industries found that the aviation industry's NTS taxonomy and behavioural markers would make a good starting point for developing a system for use in the maritime industry. The taxonomy and behavioural markers were presented to each expert interviewee for their feedback.

The initial taxonomy and behavioural marker systems had 26 elements and 4 categories. Based on the experts' opinion during the interviews and since some elements such as "conflict resolution" were non-observable; 6 elements out of 26 elements were removed from the system to be applied.

The behavioural markers to be utilised in the assessment of deck officers' NTS were formed in to a framework for ease of use in the observation stage of the study. As an example, the decision making NTS and its related behavioural markers are shown in Table 3. There are five levels of performance in this behavioural marker system. These range from very good practice to very poor practice. By using these behavioural markers an examiner is able to rate a student's performance in a ship's bridge simulator.

5 BRIDGE SIMULATOR STUDY (STEP 3)

The main aim of the bridge simulator study was to develop a method which could quantitatively assess NTS of the deck officers in a bridge simulator environment. For conducting this study a set of volunteer students were recruited to take part. The participants were volunteer students who have completed their course of study for Chief Mates certificate of Competency. LJMU ethical approval was obtained for the study and students' content was obtained.

The simulator performance was observed by the main researcher of this study, Dr Farhan Saeed who is master mariner with ten years seagoing experience and fourteen years teaching and training experience to deck officers. During the simulator observation, the researcher observed and rated participants' performance against the behaviour marker assessment framework (Table 4, 5, 6, and 7).

5.1 *Bridge simulator scenario*

The following scenario was developed for the assessment of NTS of merchant navy deck officers in a bridge simulator environment:

The vessel was alongside the jetty in Southampton. The bridge team would have to pilot their own vessel and maintain all the records as agreed by the members. Each team would need to manoeuvre their own vessel with the use of a bow thruster (team was not allowed to use tugs). There would be a number of inbound as well as outbound vessels during the departure. A grounded vessel in the vicinity of the Nab tower with a salvage operation underway would request a wide berth.

Just after passing Fawley Terminal, Gyro No. 1 would start to drift at a rate of 1°/sec. Based on the position of the vessel at the time of passing there would be the possibility of interaction with large inbound containerships.

This exercise is designed to allow participants to demonstrate their teamwork, situational awareness, leadership, and decision making skills.

Table 3. Decision making elements and behavioural markers

Element	Very Good Practice	Good Practice	Acceptable Practice	Poor Practice	Very Poor Practice
Problem definition and diagnosis	Gather all information to identify problem	Gather sufficient information to identify problem	Gather just enough information to identify problem	Gather little information to identify problem	Failure to diagnose the problem
	Review all casual factors with other crew members	Review enough casual factors with other crew members	Review some casual factors with other crew members	Review very few casual factors with other crew members	No discussion of probable cause
Option generation	States all alternative option	States enough alternative option	States some alternative option	States very few alternative option	Does not search for information
	Asks crew members for all options	Asks crew members for enough options	Asks crew members for some options	Asks crew members for very few options	Does not ask crew for alternatives
Risk Assessment and option selection	Considers and shares all estimated risk of alternative options	Considers and shares substantial shares substantial estimated risk of alternative options	Considers and shares just enough estimated risk of alternative options	Inadequate discussion of limiting factors with crew	No discussion of limiting factors with crew
	Confirms and states all selected options/ agreed action	Confirms and states enough selected options/ agreed action	Confirms and states some selected options/ agreed action	Confirms and states very few selected options/ agreed action	Does not inform crew of decision path being taken
Outcome review	Complete checking of outcome against plan	Substantial checking of outcome against plan	Average checking of outcome against plan	Little checking of outcome against plan	Fails to check selected outcome against plan

Table 4. Teamworking

Element	Very Good Practice	5	4	3	2	1	Very Poor Practice
Team building and maintaining	Fully encourages input and feedback from others			x			Keeps barriers between team members
Considering others	Take notice of the suggestions of other team members		x				Ignores suggestions of other team members
	Considers condition of other team members into account				x		Does not take account of the condition of other team members
	Provide detailed personal feedback				x		Show no reaction to other team members
Supporting others	Provide ample help to other team members in demanding situation			x			Do not help other team members in demanding situation
	Offers very good assistance			x			Does not offer assistance
Communication	Establish total atmosphere for open communication				x		Blocks open communication
	Communicates very effectively				x		Ineffective communication
Information sharing	Shares information among all team members			x			Does not share information properly among all team members

Table 5. Leadership and Managerial Skills

Element	Very Good Practice	5	4	3	2	1	Very Poor Practice
Use of Authority and assertiveness	Takes full initiative to ensure crew involvement and task completion				x		Hinders or withholds crew involvement.
	Takes full control if situation requires					x	Does not show initiative for decision
	Totally reflects on suggestions of others			x			Ignores suggestions of others
Providing and Maintaining standards	Demonstrates complete will to achieve top performance			x			Does not care for performance effectiveness.
Planning and Co-ordination	Completely encourages crew participation in planning and task completion			x			Does not encourage crew participation in planning and task completion
	Plan is well clearly stated and confirmed				x		Plan is not clearly stated and confirmed
	Well clearly states goals and boundaries for task completion				x		Goals and boundaries remain unclear
Workload	Completely notifies signs of				x		Ignores signs of fatigue

Management	stress and fatigue						
	Allots good time to complete tasks					x	Allots very little time to complete tasks
Prioritisation	Demonstrate very good prioritisation of tasks				x		Demonstrate no prioritisation of tasks
Task Delegation	Delegates all tasks properly					x	Does not delegate tasks
Initial crisis management	Identifies initial crisis situation very quickly and respond accordingly					x	Does not identify initial crisis situation

Table 6. Situation Awareness

Element	Very Good Practice	5	4	3	2	1	Very Poor Practice
Awareness of bridge systems	Fully monitors and report changes in systems' states			x			Do not monitors changes in systems' states
Awareness of external environment	Collects full information about environment (own ship's position, traffic and weather)				x		Does not collect information about environment (own ship's position, traffic and weather)
	Shares complete key information about environment with team members			x			Does not share key information about environment with crew
Awareness of time	Fully discuss time constraints with other team members				x		Does not discuss time constraints with other CM
Situation Assessment	Makes full assessment of changing situation					x	Does not make an assessment of changing situation

Table 7. Decision making

Element	Very Good Practice	5	4	3	2	1	Very Poor Practice
Problem definition and diagnosis	Gather all information to identify problem		x				Failure to diagnose the problem
	Review all casual factors with other crew members				x		No discussion of probable cause
Option generation	States all alternative option				x		Does not search for information
	Asks crew members for all options					x	Does not ask crew for alternatives
Risk Assessment and option selection	Considers and shares all estimated risk of alternative options				x		No discussion of limiting factors with crew
	Confirms and states all selected options/agreed action			x			Does not inform crew of decision path being taken
Outcome review	Complete checking of outcome against plan					x	Fails to check selected outcome against plan

6 NTS ASSESSMENT OF DECK OFFICER IN A BRIDGE SIMULATOR (STEP 4)

The following is a rundown of the participants' performance during the scenario established in step 3. They were rated against their performance on the developed behavioural markers assessment framework (Table 4, 5, 6, and 7).

The passage plan was already prepared a day before the exercise. The group tested all bridge equipment and completed the check lists. The exercise started when the bridge team was ready. Initially they had some doubts about departing the berth without tugs. The use of the bow thruster helped them to depart without any problems. The vessel was manoeuvred slowly and left the berth and headed towards the channel. The vessel speed was 8 knots in the channel. The master was in overall command, the chief officer and OOW were performing navigation and communication duties respectively. At one point their vessel grounded and then re-floated quickly. The gyro started drifting but the bridge team considered that the vessel was drifting due to tide/current. The OOW suggested that the drifting was due to the gyro failure but the master did not investigate it further and it was assumed that the vessel was drifting due to heavy current. The master only realised the gyro failure once the large alteration of the vessel's course was observed (about half an hour after the initial drift). Immediately action was taken by switching to the backup gyro and controlling the situation.

Gyro failure during the exercise was the key moment and it was expected that the bridge team would identify and take corrective measures immediately. The group's poor performance was due to lack of situation awareness of the team and then the master's over reliance on the chief officer rather than taking control of the situation himself.

The students' behaviour markers are tabulated in Table 4, 5, 6, 7. After feeding this input in to the model (Figure 1: Deck Officers' NTS Taxonomy) and using the ER algorithm, an output result set was generated as shown in Table 8 and Figure 3.

Table 8. ER results of the group performance

Very Poor	35.39%
Poor	33.71%
Average	28.05%
Good	2.85%
Very Good	0.0%

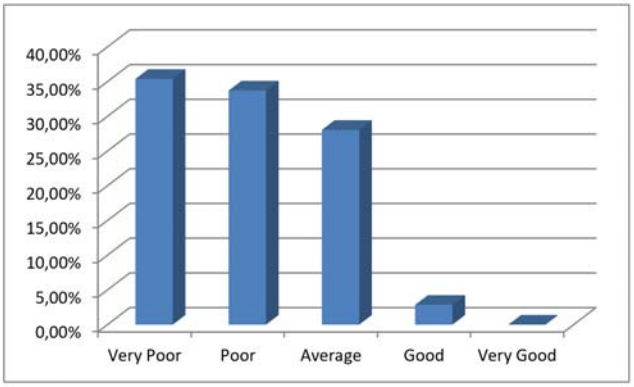

Figure 3. ER results of the group performance

7 PERFORMANCE CALCULATION BY ER ALGORITHM AND UTILITY VALUE (STEP 5)

After rating the performance of deck officers on a rating scale of 1-5 (where 5 is very good practice and 1 is very poor practice), these ratings are fed into ER formula to obtain aggregate of each scale. Utility Value is used to obtain a final value of the performance of deck officers.

The ER algorithm can be analysed and explained as follows (Riahi et al., 2012):

Let *R* represent a set with five linguistic terms (i.e. very poor, poor, average, good and very good) with their associated belief degrees (i.e. β) and be synthesised by two subsets R_1 and R_2 from two different assessments. Then, for example, *R*, R_1 and R_2 can separately be expressed by:

$$R = \{\beta^1 \text{Very Poor},\ \beta^2 \text{Poor},\ \beta^3 \text{Average},\ \beta^4 \text{Good},\ \beta^5 \text{Very Good}\}$$
$$R_1 = \{\beta_1^1 \text{Very Poor},\ \beta_1^2 \text{Poor},\ \beta_1^3 \text{Average},\ \beta_1^4 \text{Good},\ \beta_1^5 \text{Very Good}\}$$
$$R_2 = \{\beta_2^1 \text{Very Poor},\ \beta_2^2 \text{Poor},\ \beta_2^3 \text{Average},\ \beta_2^4 \text{Good},\ \beta_2^5 \text{Very Good}\}$$

Suppose that the normalised relative weights of two assessments in the evaluation process are given as w_1 and w_2 ($w_1 + w_2 = 1$) . w_1 and w_2 can be estimated by using an AHP technique. Suppose that M_1^m and M_2^m (m = 1, 2, 3, 4, 5) are individual degrees to which the subsets R_1 and R_2 support the hypothesis that the evaluation is confirmed to the five linguistic terms. Then, M_1^m and M_2^m are obtained as:

$$\begin{aligned} M_1^m &= w_1 \beta_1^m \\ M_2^m &= w_2 \beta_2^m \end{aligned} \tag{6}$$

Suppose that H_1 and H_2 are the individual remaining belief values unassigned for M_1^m and M_2^m (m = 1, 2, 3, 4, 5). Then H_1 and H_2 are expressed as:

$$\begin{aligned} H_1 &= \bar{H}_1 + \tilde{H}_1 \\ H_2 &= \bar{H}_2 + \tilde{H}_2 \end{aligned} \tag{7}$$

where $\bar{H}_n$ (n = 1, 2) represent the degree to which the other assessor can play a role in the assessment, and $\tilde{H}_n$ (n = 1, 2) is caused by the possible incompleteness in the subsets R_1 and R_2. $\bar{H}_n$ (n = 1 or 2) and H_n (n = 1, 2) are described as:

$$\begin{aligned} \bar{H}_1 &= 1 - w_1 = w_2 \\ \bar{H}_2 &= 1 - w_2 = w_1 \\ \tilde{H}_1 &= w_1 \left(1 - \sum_{m=1}^{5} \beta_1^m\right) \\ \tilde{H}_2 &= w_2 \left(1 - \sum_{m=1}^{5} \beta_2^m\right) \end{aligned} \tag{8}$$

Suppose that $\beta^{m'}$ (m = 1, 2, 3, 4 or 5) represents the non-normalised degree to which the reliability evaluation is confirmed to each of the five linguistic terms as a result of the synthesis of the judgements produced by assessors 1 and 2. Suppose that H'_U represents the non-normalised remaining belief unassigned after the commitment of belief to the five linguistic terms because of the synthesis of the judgements produced by assessors 1 and 2. The ER algorithm is stated as:

$$\begin{aligned} \beta^{m'} &= K\left(M_1^m M_2^m + M_1^m H_2 + M_2^m H_1\right) \\ \overline{H'}_U &= K\left(\bar{H}_1 \bar{H}_2\right) \\ \widetilde{H'}_U &= K\left(\tilde{H}_1 \tilde{H}_2 + \tilde{H}_1 \bar{H}_2 + \tilde{H}_2 \bar{H}_1\right) \\ K &= \left(1 - \sum_{T=1}^{5} \sum_{\substack{R=1 \\ R \neq T}}^{5} M_1^T M_2^R\right)^{-1} \end{aligned} \tag{9}$$

After the above aggregation, the combined degrees of belief are generated by assigning H'_U back to five linguistic terms using the normalisation process:

$$\begin{aligned} \beta^m &= \frac{\beta^{m'}}{1 - \overline{H'}_U} \qquad (m = 1, 2, 3, 4, 5) \\ H_U &= \frac{\widetilde{H'}_U}{1 - \overline{H'}_U} \end{aligned} \tag{10}$$

where, H_U is the unassigned degree of belief representing the extent of incompleteness in the overall assessment. The above gives the process of combining two subsets. If three subsets are required to be combined, the result obtained from the combination of any two subsets can be further synthesised with the third subset using the above algorithm. In a similar way, the judgements of multiple assessors of lower-level criteria in the chain system (i.e. components or subsystems) can be combined.

As an example, based on the ER algorithm two quantitative data (e.g. *R1 and R2)* are aggregated as follows:

– *R1* stands for 'Problem definition and diagnosis' (sub criteria of decision making) assessed for a team performance (Table 7 and 9).

– R_2 stands for 'Option generation' (sub criteria of decision making) assessed for a team performance (Table 7 and 9).

Table 9: Sub Criteria for decision making

	R_1	R_2
Very Poor	0	0.5
Poor	0.5	0.5
Average	0	0
Good	0.5	0
Very Good	0	0
Weight (w_n)	0.2447	0.2069

$w_1 + w_2 = 0.2447 + 0.2069 = 0.4516$

Normalised weights $w_1 = 0.2447 \times 2.21435 = 0.54185$

Normalised weights $w_2 = 0.2069 \times 2.21435 = 0.45815$

$$\beta_1^1 = 0,\ \beta_1^2 = 0.5,\ \beta_1^3 = 0,\ \beta_1^4 = 0.5,\ \beta_1^5 = 0$$
$$\beta_2^1 = 0.5,\ \beta_2^2 = 0.5,\ \beta_2^3 = 0,\ \beta_2^4 = 0,\ \beta_2^5 = 0$$

$$M_1^1 = w_1\beta_1^1 = 0.54185 \times 0 = 0$$
$$M_1^2 = w_1\beta_1^2 = 0.54185 \times 0.5 = 0.27093$$
$$M_1^3 = w_1\beta_1^3 = 0.54185 \times 0 = 0$$
$$M_1^4 = w_1\beta_1^4 = 0.54185 \times 0.5 = 0.27093$$
$$M_1^5 = w_1\beta_1^5 = 0.54185 \times 0 = 0$$
$$M_2^1 = w_2\beta_2^1 = 0.45815 \times 0.5 = 0.22908$$
$$M_2^2 = w_2\beta_2^2 = 0.45815 \times 0.5 = 0.22908$$
$$M_2^3 = w_2\beta_2^3 = 0.45815 \times 0 = 0$$
$$M_2^4 = w_2\beta_2^4 = 0.45815 \times 0 = 0$$
$$M_2^5 = w_2\beta_2^5 = 0.45815 \times 0 = 0$$

$$\bar{H}_1 = 1 - w_1 = 1 - 0.54185 = 0.45815$$
$$\bar{H}_2 = 1 - w_2 = 1 - 0.45815 = 0.54185$$

$$\tilde{H}_1 = w_1\left(1 - \left(\beta_1^1 + \beta_1^2 + \beta_1^3 + \beta_1^4 + \beta_1^5\right)\right) =$$
$$0.54185\left(1 - \left(0 + 0.5 + 0 + 0.5 + 0\right)\right) = 0$$
$$\tilde{H}_2 = w_2\left(1 - \left(\beta_2^1 + \beta_2^2 + \beta_2^3 + \beta_2^4 + \beta_2^5\right)\right) =$$
$$0.45815\left(1 - \left(0.5 + 0.5 + 0 + 0 + 0\right)\right) = 0$$
$$H_1 = \bar{H}_1 + \tilde{H}_1 = 0.45815 + 0 = 0.45815$$
$$H_2 = \bar{H}_2 + \tilde{H}_2 = 0.54185 + 0 = 0.54185$$

$$K = (1 - \sum_{T=1}^{5}\sum_{\substack{R=1 \\ R \neq T}}^{5} M_1^T M_2^R)^{-1}$$

$$K = \left(1 - \sum_{T=1}^{5}(M_1^T M_2^1 + M_1^T M_2^2 + M_1^T M_2^3 + M_1^T M_2^4 + M_1^T M_2^5)\right)^{-1}$$

$$K = \left(1 - \begin{bmatrix} (M_1^1 M_2^1 + M_1^1 M_2^2 + M_1^1 M_2^3 + M_1^1 M_2^4 + M_1^1 M_2^5) + \\ (M_1^2 M_2^1 + M_1^2 M_2^2 + M_1^2 M_2^3 + M_1^2 M_2^4 + M_1^2 M_2^5) + \\ (M_1^3 M_2^1 + M_1^3 M_2^2 + M_1^3 M_2^3 + M_1^3 M_2^4 + M_1^3 M_2^5) + \\ (M_1^4 M_2^1 + M_1^4 M_2^2 + M_1^4 M_2^3 + M_1^4 M_2^4 + M_1^4 M_2^5) + \\ (M_1^5 M_2^1 + M_1^5 M_2^2 + M_1^5 M_2^3 + M_1^5 M_2^4 + M_1^5 M_2^5) \end{bmatrix}\right)^{-1}$$

$$K = 1.2288$$

$$\bar{H}_{U'} = K\left(\bar{H}_1\bar{H}_2\right) = 0.3050$$
$$B^{1'} = K\left(M_1^1 M_2^1 + M_1^1 H_2 + M_2^1 H_1\right) = 0.1289$$
$$\beta^1 = \frac{B^{1'}}{1 - \bar{H}_{U'}} = 0.18547$$
$$B^{2'} = K\left(M_1^2 M_2^2 + M_1^2 H_2 + M_2^2 H_1\right) = 0.3857$$
$$\beta^2 = \frac{B^{2'}}{1 - \bar{H}_{U'}} = 0.55496$$
$$B^{3'} = K\left(M_1^3 M_2^3 + M_1^3 H_2 + M_2^3 H_1\right) = 0$$
$$\beta^3 = \frac{B^{3'}}{1 - \bar{H}_{U'}} = 0$$
$$B^{4'} = K\left(M_1^4 M_2^4 + M_1^4 H_2 + M_2^4 H_1\right) = 0.1805$$
$$\beta^4 = \frac{B^{4'}}{1 - \bar{H}_{U'}} = 0.25971$$
$$B^{5'} = K\left(M_1^5 M_2^5 + M_1^5 H_2 + M_2^5 H_1\right) = 0$$
$$\beta^5 = \frac{B^{5'}}{1 - \bar{H}_{U'}} = 0$$

The following result was obtained from the above calculations:

	$R_{12} = R_1 \oplus R_2$
Very Poor	18.547%
Poor	55.496%
Average	0
Good	25.971%
Very Good	0

The calculation is repeated for R_3 and R_4 and then again repeated to aggregate the R_{12} (i.e. $R_1 \oplus R_2$) and R_{34} (i.e. $R_3 \oplus R_4$) to find the final value of the 'decision making' element of the group.

7.1 *Obtaining Utility Value*

The main aim of using a utility approach was to obtain a single crisp number for the top-level criterion (the final result or goal) of each alternate in order to rank them. Let the utility of an evaluation grade H_n be

denoted by $u(H_n)$ and $u(H_{n+1}) > u(H_n)$ if H_{n+1} is preferred to H_n; $u(H_n)$ can be estimated using the decision marker's preferences. If no preference information is available, it could be assumed that the utilities of evaluation grades are equidistantly distributed in a normalised utility space. The utilities of evaluation grades that are equidistantly distributed in a normalised utility space are calculated as

$$u(H_n) = \frac{V_n - V_{min}}{V_{max} - V_{min}} \quad (11)$$

where V_n is the ranking value of the linguistic term H_n that has been considered, V_{max} is the ranking value of the most-preferred linguistic term H_N *and* V_{min} is the ranking value of the least-preferred linguistic term H_l.

The utility of the top level or general criterion S€ is denoted by $u(S(E))$. If $\beta_H \neq 0$ (i.e. the assessment is incomplete, $\beta_H = 1 - \sum_{n=1}^{N} \beta_n$) there is belief interval $[\beta_n, (\beta_n + \beta_H)]$, which provides likelihood that $S(E)$ is assessed to H_n. Without loss of generality, suppose that the least-preferred linguistic term having the lowest utility is denoted by $u(H_l)$ and the most-preferred linguistic term having the highest utility is denoted by $u(H_N)$. Then the minimum, maximum and average utilities are defined as follows respectively (Riahi et al., 2012);

$$u_{min}(S(E)) = \sum_{N=2}^{N} \beta_n u(H_n) + (\beta_l + \beta_H) u(H_l)$$

$$u_{max}(S(E)) = \sum_{n=1}^{N-1} \beta_n u(H_n) + (\beta_N + \beta_H) u(H_N)$$

$$u_{average}(S(E)) = \frac{u_{min}(S(E)) + u_{max}(S(E))}{2} \quad (12)$$

Obviously if all the assessments are complete, then $\beta_H = 0$ and the maximum, minimum and average utilities of S(E) will be the same. Therefore, $u(S(E))$ can be calculated as

$$u(S(E)) = \sum_{n=1}^{N} \beta_n u(H_n) \quad (13)$$

The above utilities are used only for characterising an assessment and not for criteria aggregation.

First $u(H_n)$ values were calculated for belief values (Very Good = 5, Good = 4, Average = 3, Poor = 2, Very Poor = 1)

$$u(H_n) = \frac{V_n - V_{min}}{V_{max} - V_{min}}$$

$$u(H_5) = \frac{5-1}{5-1} = 1$$

$$u(H_4) = \frac{4-1}{5-1} = \frac{3}{4} = 0.75$$

$$u(H_3) = \frac{3-1}{5-1} = \frac{2}{4} = 0.5$$

$$u(H_2) = \frac{2-1}{5-1} = \frac{1}{4} = 0.25$$

$$u(H_1) = \frac{1-1}{5-1} = 0$$

Following Group's ER algorithm output values were used for the example calculations;

$\beta_1 = 0.3539$
$\beta_2 = 0.3371$
$\beta_3 = 0.2805$
$\beta_4 = 0.0285$
$\beta_5 = 0.000$
Total 1.000

If $\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 = 1$ then following equation will be used;

$$u(S(E)) = \sum_{n=1}^{N} \beta_n u(H_n)$$

$$u(S(E)) = \beta_1 u(H_1) + \beta_2 u(H_2) + \beta_3 u(H_3) + \beta_4 u(H_4) + \beta_5 u(H_5)$$

$$u(S(E)) = 0.2459$$

8 RESULT AND DISCUSSION

Deck officers' NTS taxonomy was developed using interviews and AHP, which provided the weights of the each skill and element. These weights were fed into ER algorithm while aggregating participants' NTS performance in a bridge simulator environment.

The examiner observed the students' NTS in a ship's bridge simulator by using behavioural markers, the assessment data was then aggregated using the ER algorithm. As part of the ER calculations, a utility value was obtained for the group's NTS, which provided a crisp number. The final group performance value was found to be 24.59%.

24.59% is a poor result. Unfortunately, the discussion on how to improve a deck officer's performance in a crisis situation is outside the focus of this paper. Further research may be required to address this issue. What is important here is that this method has made it possible to quantitatively assess the NTS performance of merchant navy deck officers in a bridge simulator and provide a crisp number.

Assessing students can be an intensive process for an examiner. It would be completely unrealistic to expect an examiner to perform the calculations for observations on each criteria at the same time as observing students' performance. To overcome this difficulty, the Intelligent Decision System for Multiple Criteria Assessment software was used. It is expected that this would also be the case with future assessments. Observed values were entered in to the software to get a result quickly. In this case, to prove the reliability of the results generated by this software

the results were tested against manual calculations (Section 7.0) and found to be accurate.

9 CONCLUSION

This methodology has now made it possible to quantitatively assess the NTS of deck officers in a bridge simulator. The necessary calculations can be performed by the Intelligent Decision System for Multiple Criteria Assessment software as the examiner may not have the skill set or time to perform the calculations for each observation. The use of the software makes it easy to input the values and obtain the final results in a timely fashion.

ACKNOWLEDGEMENT

The material and data in this publication have been obtained through the funding and support of the International Association of Maritime Universities (IAMU) and The Nippon Foundation in Japan.

REFERENCES

Aull-Hyde, R., Erdogan, S. and Duke, J. D. (2006) An experiment on the consistency of aggregated comparison matrices in AHP. *European journal of operational research*, 171, pp. 290-295.

Balci M. B. C., Tas, T., Hazar, A. I., Aydin, M., Onuk, O., Cakiroglu, B., Fikri, O., Ozkan, A. and Nuhoglu, B. (2014) Applicability and effectiveness of virtual reality simulator training in urology surgery: A double-blind randomised study. *Noble medicus* 29, 10 (2), pp. 66-71.

Coyle, G. (2004) *The Analytic Hierarchy Process (AHP).* Practical Strategy. Open Access Material. AHP. Pearson Education Limited.

Fletcher, G., Flin, R and Mcgeorge, P. (2003b) Interview study to identify anaesthetists' non-technical skills. *University of Aberdeen* SCPMDE Project: RDNES/991/C.

Flin, R., Martin, L., Geosters, K., Hoermann, J., Amalberti, R., Valot, C., and Nijhuis, H. (2003) Development of the NOTECHS (Non-Technical Skills) system for assessing pilots' CRM skills. *Human Factors and Aerospace Safety*, 3 (2), pp. 95-117.

Gatfield D., (2008) *Behavioural markers for the assessment of competence in crisis management*. PhD thesis, Southampton Solent University.

Helmreich, R. L., Merritt, A. C., and Wilhelm, J. A. (1999) The evolution of Crew Resource Management Training in commercial aviation. *The International Journal of Aviation Psychology*, 9(1), pp. 19-32.

Ishizaka, A. and Labib, A. (2009) Analytic Hierarchy Process and Expert Choice: Benefits and Limitation, *OR Insight*, 22(4), p201-220.

Klampfer, B., Flin, R. Helmreich R., Hausler, R., Fletcher, G., Field, P., Staender, S., Lauche, K., Dieckmann, A. and Amacher, A. (2001) Behavioural Markers Workshop. Group interaction in high risk environment (GIHRE) project. *GIHRE-Aviation: Swiss Federal Institute of Technology (ETH) Zurich*, Swiss training centre, 5-6 July 2001.

Kozuba, J. and Bondaruk, A. (2014) Flight simulator as an essential device supporting the process of shaping pilot's situational awareness. *International conference of scientific paper*, AFASES 2014, Brasor, 22-24 May 2014, pp. 695-714.

Micheal, M., Abboudi, H., Ker, J., Khan, M. S., Dasgupta, P. and Ahmed, K. (2014) Performance of technology-driven simulators for medical students – A systemic review. *Journal of surgical research*, 192, pp. 531-543.

Mitchell L., Flin R., Yule S., Mitchell J., Coutts K. and Youngson G. (2013) Development of a behavioural marker system for scrub practitioners' non-technical skills (SPLINTS system). *Journal of evaluation in clinical practice*, 19, pp.317-323.

Mohovic, R., Rudan, I. and Mohovic, D. (2012) Problems during simulator training in ship handling education. *Scientific Journal of Maritime Research*, 26 (1), pp.191-199.

Pelletier, S. (2006) The role of navigation simulator technology in marine pilotage. *International Maritime Pilotage Association 18th Congress*, Havana, Cuba, 23rd November 2006, pp. 1-5.

Riahi, R., Bonsall, S., Jenkinson, I. and Wang, J. (2012) A Seafarer's reliability assessment incorporating subjective judgements. *Journal of Engineering for the maritime environment*, 226 (4), pp. 313-334.

Saaty T. L. (1990) How to make decisions: The Analytic Hierarchy Process. *European Journal of operational Research*, 48 (1): pp. 9-26.

Sniegocki, H. (2005) Impact of the usage of visual simulator on the students training results. Conference paper, *International Conference on modelling and simulation general application and models in engineering science*, Gdynia Maritime University.

Wall, A. D. (2015), Subject Head, LJMU, Interview, 27th May, 2015.

Wanger, R., Razek, V., Grafe, F., Berlarge, T., Janousek, J., Daehnert, I., and Weidenbach, M. (2013), Effectiveness of simulator-based echocardiography training of non-cardiologists in congenital heart diseases. *Echocardiography*, Wiley periodicals, Inc.DOI:10.1111/echo. 12118, pp. 693-698.

Winter, J. C. F, Dodou, D. and Mulder, M (2012) Training effectiveness of whole body flight simulator motion: A comprehensive EMta-Analysis. *The International Journal of Aviation Psychology*, 22(2), pp. 164-183.