A Validation Data-Set and Suggested Validation Protocol for Ship Evacuation Models

EDWIN R. GALEA, STEVEN J. DEERE, ROBERT BROWN and LAZAROS FILIPPIDIS
Fire Safety Engineering Group
University of Greenwich
London SE10 9LS UK

ABSTRACT

An evacuation model validation data-set collected as part of the EU FP7 project SAFEGUARD is presented. The data was collected from a cruise ship operated by Royal Caribbean International (CS). The trial was a semi-unannounced assembly trial conducted at sea and involved some 2500 passengers. The trial took place at an unspecified time however, passengers were aware that on their voyage an assembly exercise would take place. The validation data-set consists of passenger; response times, starting locations, end locations and arrival times in the assembly stations. The validation data were collected using a novel data acquisition system consisting of ship-mounted beacons, each emitting unique Infra-Red (IR) signals and IR data logging tags worn by each passenger. The results from blind simulations using maritimeEXODUS for the assembly trial are presented and compared with the measured data. Three objective measures are proposed to assess the goodness of fit between the predicted model data and the measured data.

KEYWORDS: human behavior, human factors, egress, modelling, validation, transportation

INTRODUCTION

In 2002 the International Maritime Organisation (IMO) introduced guidelines for undertaking full-scale evacuation analysis of large passenger ships using ship evacuation models [1]. These guidelines, known as IMO MSC Circular 1033, were to be used to certify that passenger ship design was appropriate for full-scale evacuation. As part of these guidelines it was identified that appropriate full-scale ship based evacuation validation data was not available to assess the suitability of ship evacuation models. As suitable validation data was not available, a series of test cases were developed which verified the capability of proposed ship evacuation software tools in undertaking simple simulations. However, these verification cases were not based on experimental data. Furthermore, successfully undertaking these verification cases does not imply that the evacuation model is validated or capable of predicting real evacuation performance. In 2007, IMO MSC Circular 1238 (MSC1238) [2], a modified set of protocols for passenger ship evacuation analysis and certification were released however, the issue of validation of passenger ship evacuation models was not addressed. The IMO Fire Protection (FP) Sub-Committee in their modification of MSC Circ. 1033 at the FP51 meeting in February 2007 [3] invited member governments to provide, "…further information on additional scenarios for evacuation analysis and full scale data to be used for validation and calibration purposes of the draft revised interim guideline." The EU framework 7 project SAFEGUARD (see http://www.safeguardproject.info/) aims to address this requirement by providing full-scale data for calibration and validation of ship based evacuation models.

As part of project SAFEGUARD, a series of five semi-unannounced full-scale assemblies were conducted at sea on three different types of passenger vessel. From these trials five passenger response time data-sets were collected and two full-scale validation data-sets. The response time data-sets have been presented in another publication [4] and the first of the validation data-sets has been presented in another paper [5]. In this paper we present the second and more comprehensive of the two Safeguard Validation Data-Sets (SGVDS) [6]. This data was generated from an assembly trial conducted on a Cruise Ship (CS) operated by Royal Caribbean International (SGVDS2).

The Royal Caribbean vessel can carry approximately 2500 passengers and 842 crew. The vessel performs several cruise holidays in the Caribbean and the Baltic Sea. Data was collected on the vessel while it was cruising in the Baltic Sea at the end of July 2010, with the assembly trial being performed on the first leg of the vessel's journey, between Harwich in the UK and Copenhagen in Denmark. The trial took place at an unspecified time however passengers were aware that an assembly exercise would take place during the first leg of the trip. The trial was undertaken during the morning on the day after the ship left Harwich and involved some 2292 passengers. The ship's alarm was sounded towards the end of breakfast and

passengers, with the help of the crew, moved to their assigned assembly stations. Each passenger was designated an assembly station, which was indicated to them on their key card (that provided access to their cabins). The data collected during the assembly trial consisted of passenger; response time data, starting locations, arrival time at the designated assembly stations and the paths taken. Some 30 digital video cameras were used to collect the response time data. The other validation data was collected using a novel data acquisition system consisting of ship-mounted beacons, each emitting unique Infra-Red (IR) signals and IR data logging tags worn by each passenger [6]. Some 106 video cameras were used to capture the response times of passengers. These included the ship's CCTV system (94 cameras) and specially installed digital video cameras (12 cameras). Given the larger size of this ship, a total of 70 IR beacons were installed and 1950 tags were worn by passengers.

The ship validation data-sets that were generated are unique for a number of reasons. Unlike most evacuation model validation data-sets, they incorporate regional information relating to the starting locations of the population in addition to the actual response time distribution for the population. Most evacuation validation data-sets lack these essential details allowing modellers the opportunity to tune their predictions in order to obtain the best fit to the experimental results. Furthermore, the trials were conducted on a real ship, at sea and were semi-unannounced making the results relevant, credible and realistic. In addition, as the start and end locations for the population are known, it is also possible to utilise the data-set to evaluate the capabilities of evacuation model route planning and wayfinding algorithms. Finally, the two data-sets represent the first comprehensive ship evacuation model validation data-sets collected.

In this paper we present SGVDS2, the blind results from the maritimeEXODUS [7-13] simulation of the validation data-set and an assessment of the level of agreement between model predictions and trial data. All the data required to define the validation data-set, including the geometry of the vessel is publicly available and can be found on the FSEG website, http://fseg.gre.ac.uk/validation/ship_evacuation.

THE SHIP GEOMETRY

The CS used in this study consists of 13 decks, of which seven decks are accommodation space consisting of passenger cabins. The other five decks consist of general circulation and entertainment spaces such as; restaurants, bars, disco, swimming pools, casino, theatre, cinema, spa/health centre, business centre, leisure pursuits (such as gymnasium, climbing wall, crazy golf, cards room) and retail areas. A CAD file was provided (in .dxf format) to define the layout of the ship within the evacuation model. The ship has 18 Assembly Stations (AS) distributed over two decks, Deck 5 and 6, of which 10 are external and eight are internal. The 10 external AS, AS b to AS f and AS r to AS v are located on Deck 5. For the purposes of the validation modelling, these are grouped together and identified as AS B and C, with AS B representing the actual AS v to r and AS C representing the actual AS b to f. AS B has three entrances located near the atrium amidships, in the shopping mall and just outside the theatre at the fore end of the vessel, while AS C has two entrances located outside the theatre in the fore of the vessel and the other located near the atrium amidships. Note that the grouped AS are actually all part of a single space.

The eight internal AS are located on Deck 5, AS a, and on Deck 6, AS g and AS w. These AS are located in the theatre (AS a) and restaurant areas (AS g and AS W). Once again, for the purposes of the validation modelling, these are grouped together and identified as AS A and D, with AS A representing the actual AS a, and AS D representing the actual AS g and AS w. AS A has two entrances located at the entrance to the theatre, while AS D has two entrances located at the atrium (amidships) and from the bar area at the aft of the vessel.

The vessel has seven main vertical zones however only three main vertical passenger staircases were available in the trial. Full details of the vessel layout may be found at: http://bit.ly/1eGeYEa.

THE SIMULATION SOFTWARE

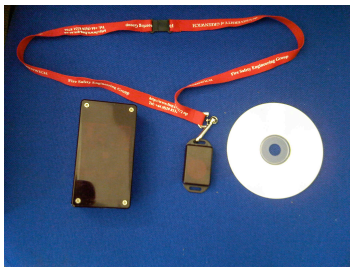
The ship evacuation simulation software maritimeEXODUS [7-13], produced by the Fire Safety Engineering Group (FSEG) of the University of Greenwich was used to perform the evacuation simulations presented in this paper. The software has been described in detail in many publications [7-13] and so only a brief description of the software will be presented here. EXODUS is suite of software to simulate the evacuation and circulation of large numbers of people within a variety of complex enclosures. maritimeEXODUS is the ship version of the software. The software takes into consideration people-people, people-fire and people-structure interactions. It comprises five core interacting sub-models; the Passenger,

Movement, Behaviour, Toxicity and Hazard sub-models. The software describing these sub-models is rule-based, the progressive motion and behaviour of each individual being determined by a set of heuristics or rules. Many of the rules are stochastic in nature and thus if a simulation is repeated without any change in its input parameters, a slightly different set of results will be generated. It is therefore necessary to run the software a number of times as part of any analysis. The submodels operate on a region of space defined by the geometry of the enclosure. The geometry can be specified automatically using a DXF file produced by a CAD package or manually using the interactive tools provided. In addition to the representation of the structure itself, the abandonment system can also be explicitly represented within the model, enabling components of the abandonment system to be modelled individually.

The software has a number of unique features such as the ability to incorporate the effects of fire products (e.g. heat, smoke, toxic and irritant gases) on crew and passengers and the ability to include the impact of heel and trim on passenger and crew performance. The software also has the capability to represent the performance of both crew and passengers in the operation of watertight doors, vertical ladders, hatches and 60 degree stairs. Another feature of the software is the ability to assign passengers and crew a list of tasks to perform. This feature can be used when simulating emergency or normal operating conditions. The version of the software used for this analysis was V5.0 beta (which is now on general release).

THE INFRA-RED TRACKING SYSTEM

The system used to track and time the movement of passengers from their starting locations to their assigned AS was an Infra-Red (IR) system based on the TagMobile system developed by the RFID Centre Ltd. The RFID Centre worked with FSEG to modify this system to make it more appropriate for use in evacuation research applications [6]. The system deployed consisted of a number of IR Beacons strategically located throughout the vessel, and IR data logging tags worn by each passenger (see Fig.1a). Each beacon generates a unique IR light field. As a tagged individual passes through the IR field, IR light sensors in the tag detect the IR light and log its ID and the time at which it was detected in the tag's internal memory. Following a trial, all the tags must be retrieved in order to determine the occupant's route data. The IR beacons are strategically placed at the main locations where passengers congregate and at the entrances to each of the AS (see Fig.1b). In this way, the initial location of each tagged passenger can be determined, which AS they go to and at what time they enter the AS. Testing of the IR tracking system demonstrated that the system was able to identify the number of passengers passing a point, even in very large crowds and record the time at which they passed the measuring point [6].



(a)

(b)

Fig.1. IR beacon and tag (a) and installing IR beacon at the entrance to an external AS (b)

To test the accuracy of the arrival times derived from the IR system, video cameras were installed at the entrance to several of the AS on the CS. This enabled a comparison of the arrival time derived from the IR system with the arrival times manually determined from the video camera record. In addition, this analysis allowed for a comparison of the total number of passengers passing through the entrance to the AS as counted by the IR system with the actual number that could be seen in the video record. The comparison was carried-out for two locations, both on the ship's starboard side of Deck 5 – one forward and one near midships. The forward location (at Beacon location 73, Camera UOG12) was a doorway with a vestibule leading to AS B. The location near midships (at Beacon 50, Camera UOG10) was a doorway that opened

directly into the same external AS. These two locations were selected as they represented examples of locations in which the beacons were expected to perform well i.e. location 50 and those which would pose a challenge for the beacons i.e. location 73. Results of the comparisons at Beacon location 73 are provided in Fig.2.

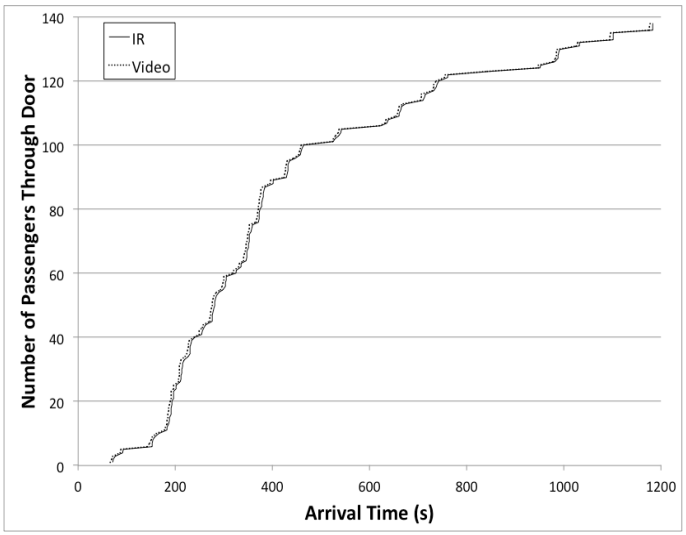


Fig.2. Comparison of passenger arrival times at Beacon 73 (Camera UOG12)

When analysing the video for both of the above locations, the time at which a passenger passed across the door line was taken as their entry time. Because a comparison was being made to the IR data, times were recorded only for passengers that could be clearly seen wearing or holding an IR tag. In addition, because of the way the IR tag data was analysed, the entry times were recorded only for passengers who entered the assembly station and remained there.

It is clearly seen that the IR data collection system matched quite closely with the data manually derived from the video record. The IR system correctly counted the number of passengers through the door (138) and timing results consistently lagged the camera results by 5.0 s on average with a standard deviation of 1.11 s (maximum difference was 10 s and minimum difference was 2 s). It is noted that the IR system accurately counted the number of passengers even in the high density situation encountered at this location. Thus the IR measured times can on average be up to 5.0 s behind the actual measured time as derived from the video data.

These results suggest that the IR system provides an accurate measure of the arrival times for passengers when compared against a synchronised video system, despite a small lag between the actual arrival time and what the IR data collection system actually measures.

THE INITIAL POPULATION DISTRIBUTION

The initial distribution of the population was determined through the use of the IR tracking system. Using the IR tag information, the initial location of each tagged passenger was determined using data related to the first IR field that the tagged passengers passed through. Using this information the initial location of each tagged passenger can be confined to a region of space on a deck defined by the IR beacons. Typical regions correspond to the physical compartments on the ship, so a region may be a restaurant, or bar area, or communal seating area. The starting deck for the 1779 tagged passengers on the CS is shown in Table 1. Also presented in Table 1 are the final AS that passengers starting on each deck ended up in.

Table 1: Starting deck location and final assembly station for each passenger in the CS trial

Deck	2	3	4	5	6	7	8	9	10	11	12	13	Total
AS A	2	26	59	26	13	33	31	25	12	153	22	0	402
AS B	1	56	101	45	7	36	38	42	40	178	25	6	575
AS C	12	41	71	31	11	35	27	29	24	133	21	2	437
AS D	4	10	52	25	21	35	30	57	28	81	22	0	365
Total	19	133	283	127	52	139	126	153	104	545	90	8	1779

THE TRIAL RESULTS

The main results for the assembly trial conducted on the CS are presented and discussed. These concern the measured assembly time for each of the tagged passengers and response time distribution associated with each starting region.

Final locations of tagged passengers at the end of the assembly trial

Of the 2500 passengers on board the CS, 1779 wore tags and so were tracked throughout the trial. Presented in Table 1 are the locations of the tagged passengers on completion of the assembly trial. For example, 402 tagged passengers ended up in AS A, of which two came from Deck 2, 26 came from Deck 3, 59 came from Deck 4. The starting region for each passenger is also known.

Passenger response time distribution

The passenger response time distribution was determined from data collected from the 106 digital video cameras located throughout the vessel. The response times for 1228 passengers were determined producing an overall response time distribution which is presented in Fig.3. The response time data-set was fitted with a log normal curve, with the following key parameters; the minimum and maximum response times are 0 s and 1379 s, while the log of the mean response time is 5.012 and the log of the standard deviation is 0.89. Since response times were not collected for all the passengers in all the various regions of the ship (due to its size) the overall response time distribution is used for SGVDS2.

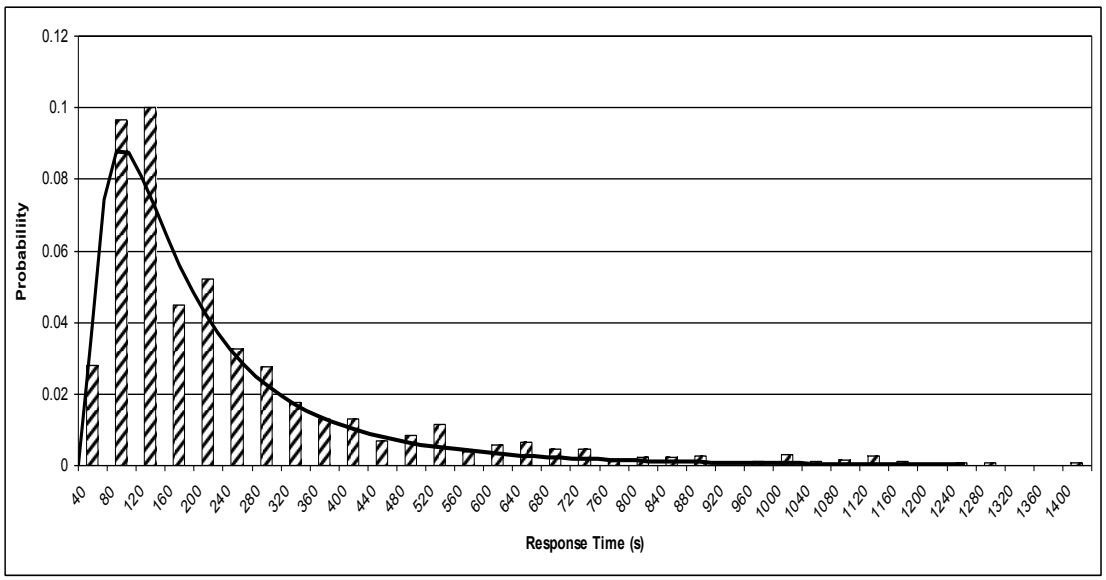


Fig.3**.** Overall log normal response time distribution for the CS assembly trial

Assembly times

The Captain officially ended the assembly exercise 29 minutes after its start. The IR data suggests that the last tagged passenger arrived in AS A after 1637 s (27 min 17 s). The arrival curves for each AS and the overall arrival curve, generated using the IR data, is presented in Fig.4. In principle, this data-set is ideal for validation purposes, as the starting locations and response times of the population is known. This means that it should be possible to remove most of the uncertainty associated with input parameters associated with response time and starting location. However, there are several complications associated with the validation data-set which introduces some degree of uncertainty in the trial results.

First, of the 2292 passengers on board, 1950 wore the IR tags and participated in the assembly trial. Of these, 171 tagged participants were excluded from the data-set for various reasons e.g. a number of participants arrived at the AS after the trial was declared over, several participants had response times considerably longer than that measured using the video camera data, another participant took a circuitous route to the AS, such as going up stairs for several decks when they should have been going down, etc. The 342 passengers that did not have tags were; (1) children under the age of 12 who were not permitted to take part in the validation study, (2) passengers who did not take part in the trial and (3) a number of passengers who decided not to wear the IR tag or forgot to wear the IR tag while participating in the trial. The number in the latter category is believed to be small (through analysis of video footage from the entrance to the AS) and estimated to be less than 10% of the number participating who wore tags. The impact of these passengers on the overall results is expected to be small and is ignored.

Secondly, the exact starting location of the tagged participants was not known, but the region where they were located was known. Spatial regions were between 50m and 95m long; thus not knowing the precise starting location of an individual may increase/decrease their arrival time by 50-95 seconds.

Thirdly, the response time distribution is not associated with a unique individual but represents the overall response time distribution for the entire passenger set. The impact that this will have on an evacuation analysis is difficult to estimate as each time the simulation is run, a different random allocation of response times is made for all agents. Thus an agent may be allocated a very long response time in one simulation and in the next simulation may be allocated a very short response time. The error associated with the random allocation of the global response time may be minimised if the average predicted assembly time distribution is considered. However, MSC 1238 requires that the 95th percentile case is used to represent the vessel assembly performance. All of these factors must be taken into consideration when determining how well the evacuation model predicts the assembly exercise.

Modelling procedures

The bulk of the parameters used in the simulation are compliant with those specified in MSC1238 [2] with the exception of the response time distribution and the initial location of the passengers; these are determined from the trial data. For the CS simulations, the global response time data is used and the initial starting locations of the passengers as defined above. It is noted that as the population demographics used in the validation analysis are derived from MSC1238 and not the actual vessels, they may not necessarily reflect the actual population demographics of the passengers involved in the trials. This may introduce some error in the overall numerical predictions of the assembly process. Furthermore, given the starting zone that an agent is assigned to, the AS that they will go to is known. This information is also imposed on the simulations presented here. The agent will go to the correct AS as defined by the trial.

As is required by MSC1238 [2] a total of 50 repeat simulations are produced, where the starting locations of the passengers within the various starting regions are randomised. In the regulatory analysis, the 95th percentile case is selected to represent the prediction of the assembly process, with the Total Assembly Time (TAT) derived from the 95th percentile time representing the overall assembly time for the vessel. The regulations assume that evacuation models will under-predict the likely total assembly time by 25% and so require that an additional 25% safety factor is added to the predicted total assembly time. The purpose of the validation exercise is to determine how well the evacuation software predicts the overall assembly process, not simply the TAT. It is possible that a poor software tool may incorrectly predict the overall assembly process but randomly produce a reasonable prediction of the TAT.

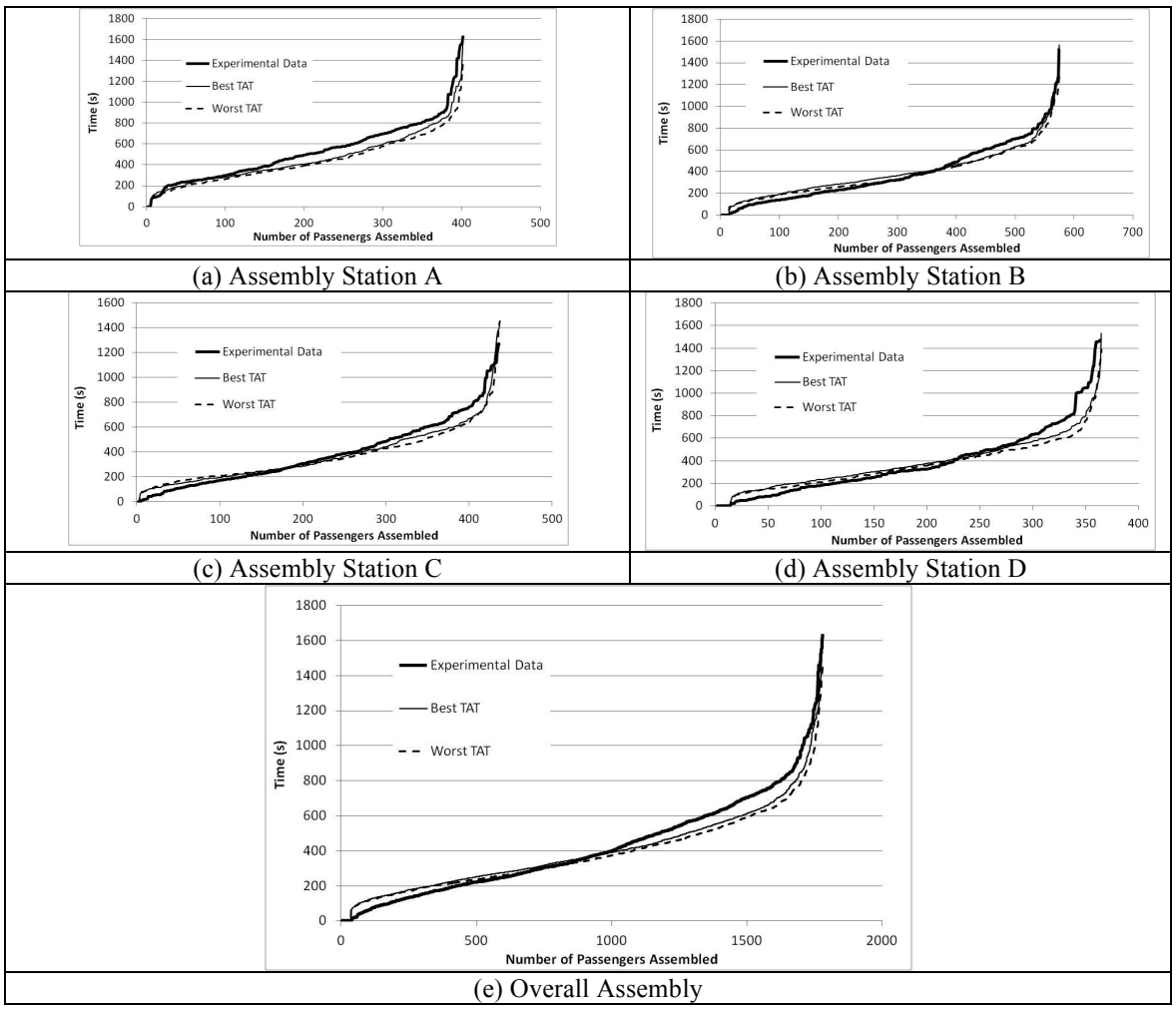
While the TAT may be the only number that the regulatory authority is concerned with, confidence in the reliability of the TAT prediction is based on how well the software predicts the overall assembly process.

Thus, the validation exercise must evaluate how well the software reproduces the overall assembly process (arrival times for each passenger) and not simply the TAT. Furthermore, just as there is a spread in the results for the numerical simulations, there would also be a spread in the experimental results if the experiment were repeated, even if all the passengers started from the same locations with the same response times as it is unlikely the passengers would do the exact same thing twice. While we have a range of numerical results for the assembly, we only have one experimental result and it is impossible to determine if the experimental result is representative of the average result for the experiment or if it is an outlier and how wide the range in experimental results is likely to be. Thus, the best numerical result will be compared with the experimental result to determine how well the software predicts the trial assembly.

From a simple visual observation of the predicted assembly curves it is difficult to identify which of the 50 curves produces the best level of agreement with the experimental results. As the regulatory authorities are primarily concerned with the prediction of the TAT, the numerical prediction producing the TAT with the smallest error is arbitrarily selected to represent the best prediction. For comparison purposes, the numerical prediction producing the TAT with the largest error is also considered. In addition, a more objective method for identifying the numerical prediction which produces the best level of agreement with the experimental data is identified later in the paper.

COMPARING MODEL PREDICTIONS WITH TRIAL RESULTS

The numerical predictions producing the TAT with the smallest and greatest error are presented in Fig.4. along with the experimental data for the CS. The measured and predicted arrival curves are presented for each AS (Fig.4a. to Fig.4d.) and the overall arrival curve (Fig.4e.).



(a) Assembly Station A

(b) Assembly Station B

(c) Assembly Station C

(d) Assembly Station D

(e) Overall Assembly

Fig.4. Comparison of model predictions with experimental data for CS

As can be seen from Fig.4, the numerical simulations under-predict the TAT for the overall assembly process and either under- (negative values) or over-predict (positive values) the assembly time for each AS. The simulation producing the best/worst TAT under- or over-predicts the TAT for each AS by; -0.1%/-16%, 2%/-8%, 12%/-0.3% and 4%/-5% respectively and the overall TAT is under-predicted by -0.1%/-14% (see Table 4). Thus the error in predicting the overall TAT is between -0.1% and -14% while the error in predicting the TAT for each assembly station varies from -16% (under-prediction) to 12% (over-prediction).

As can be seen by comparing the model predictions for the CS assembly trial (Fig.4.) with the trial results, the predictions are quite close to the experimental data. As described earlier, there are several other uncertainties introduced into the experimental data which should be considered when assessing the level of agreement between model predictions and experimental data.

The uncertainty in the exact starting location of the passengers can introduce an error of 50s to 95s in the prediction of the assembly times. This uncertainty alone introduces a possible error of some 6% in the overall TAT and an error of some 8% in the prediction of the TAT for each AS. The error associated with using the global response time distribution rather than the actual response time for an agent is difficult to estimate but may be appreciable. Finally, the error associated with the untagged passengers is expected to be small, and the 5 s measurement error in the arrival times associated with using the IR system is considered insignificant for this trial (less than 0.4% for the TAT). Taking these uncertainties into consideration, the differences in the predicted assembly times appear reasonable.

It is also noted that the numerical simulations with the best TAT (and also the 95th percentile case) correctly identifies that the last AS to assemble is AS A. Furthermore, there does not appear to be a significant difference between the predicted assembly curves for the best and worst TAT. By sight, the predicted and measured assembly curves for the overall assembly appear to be in very good agreement (see Fig.4e.). The predicted arrival curves for each of the AS (Fig.4a. to Fig.4d.) also appear to be in very good agreement with the measured curves. This suggests that the evacuation model is doing a good job of predicting the overall assembly process. Furthermore, the level of agreement with the CS data-set appears to be significantly better than that of the RP1 data-set [5].

VALIDATION METRIC

While the evacuation simulation software appears to be producing reasonable predictions of the assembly process it is desirable to have objective measures of the level of agreement between predicted and measured performance rather than subjective assessments. This is particularly important if the validation analysis is to be used by regulatory authorities to determine the suitability of an evacuation modelling tool. Thus it is necessary to quantify the level of agreement between predicted and measured performance.

In [14] several metrics are presented which can be used to quantify the level of agreement between predicted and measured values. However, the mathematical formulations presented in [14] have a number of typographical errors [15] and are here presented correctly. Before presenting the formulation of the metrics it is necessary to introduce some terminology. The series of measured experimental data is represented by the n-dimensional vector $E = (E_1, E_2, \ldots E_n)$, where E_i represents the measured assembly time for the ith passenger. Similarly, the series of predicted model data is represented by the vector $m = (m_1, m_2, \ldots m_n)$, where m_i represents the predicted assembly time for the ith agent. The metric used to quantify the level of agreement between predicted and measured values consists of three measures (see equations 1 to 3).

$$\frac{\|E-m\|}{\|E\|} = \frac{\sqrt{\sum_{i=1}^{n}(E_i-m_i)^2}}{\sqrt{\sum_{i=1}^{n}E_i^2}} \qquad (1)$$

$$\frac{\langle E,m\rangle}{\|m\|^2}=\frac{\sum_{i=1}^{n}E_i m_i}{\sum_{i=1}^{n}m_i^2} \tag{2}$$

$$\frac{\langle E,m\rangle}{\|E\|\|m\|}=\frac{\sum_{i=s+1}^{n}\frac{(E_i-E_{i-s})(m_i-m_{i-s})}{s^2(t_i-t_{i-1})}}{\sqrt{\sum_{i=s+1}^{n}\frac{(E_i-E_{i-s})^2}{s^2(t_i-t_{i-1})}\sum_{i=s+1}^{n}\frac{(m_i-m_{i-s})^2}{s^2(t_i-t_{i-1})}}} \tag{3}$$

The first is the Euclidean Relative Difference (ERD) defined by equation 1. This is used to assess the average difference between the experimental data (E_i) and the model data (m_i). This equation should return a value of 0 if the two curves are identical in magnitude. The smaller the value for the ERD, the better the overall agreement. An ERD of 0.2 suggests that the average difference between the model and experimental data points, taken over all the data points is 20%.

The second measure is the Euclidean Projection Coefficient (EPC) defined by equation 2. The EPC calculates a factor which when multiplied by each model data point (mi) reduces the distance between the model (m) and experimental (E) vectors to its minimum. Thus the EPC provides a measure of the best possible level of agreement between the model (m) and experimental (E) curves. An EPC of 1.0 suggests that the difference between the model (m) and experimental (E) vectors are as small as possible. The third measure is the Secant Cosine (SC) defined by equation 3. Unlike the other two measures, it provides a measure of how well the shape of the model data curve matches that of the experimental data curve. It makes use of the secants (which approximate to tangents) through both curves. An SC of 1.0 suggests that the shape of the model (m) curve is identical to that of the experimental (E) curve.

The t in equation 3 is a measure of the spacing of the data. For the assembly data presented in Fig.4, the spacing of the data is 1 i.e. there is a data point for each passenger/agent that enters an AS. Thus the difference in t consecutive values in equation 3 is 1. The s in equation 3 is a factor that represents the period of noise in the data, or variations in the experimental data resulting from microscopic behaviour not possible to reproduce in the model. Selecting a value of s which is greater than the period of the noise in the data provides a means to smooth out the effect of the noise. However, care must be taken in selecting the value of s. If s is too large the natural variation in the data may be lost, while if s is too small, the variation in the data created by noise may dominate the analysis. Selecting an appropriate value of s is dependent on the number of data points in the data-set, given by n. Thus it is desirable to keep the ratio s/n as low as possible.

For data-sets in which an experimental and model data point are available for each person, if the ERD = 0.0, then it would not be necessary to consider other measures as the two data-sets would be identical. In all other cases it is necessary to consider the three measures together in order to get a good indication of how well the two data-sets match each other. As the model data curve can cross the experimental data curve one or multiple times (as shown in Fig.4.) EPC can return a value close to 1.0 while there is a difference between the two curves. Similarly, the SC can return a value of 1.0 even though the model and experimental data curves are off set by a constant value. In general, for the model and experimental curves to be considered a perfect match, it is necessary to have all three measures at their optimal values i.e. ERD = 0.0, EPC = 1.0 and SC =1.0.

Validation Metric applied to maritimeEXODUS predictions of SGVDS2

If the metric is applied to the data shown in Fig.4. it produces the values presented in Table 2. First consider the data relating to the overall assembly curve for all three cases i.e. best ERD and best/worst TAT. The values for the SC suggest that the shape of the overall assembly curve closely (SC $\geq$ 0.9) resembles that of the experimental data, even with s/n as low as 0.01. This is consistent with the

conclusion drawn from a visual inspection of Fig.4e. Note that an s/n of 0.01 represents 1% of the data-set and implies s = 17 for this data-set. Thus for the 1743 point data-set, the gradients used in the evaluation of equation 3 are spread over 17 data points, which is considered reasonable. Furthermore, the ERD for the overall assembly is very low (< 0.15) and the EPC is equal to 1.1 suggesting that the overall predicted assembly curve is very close to the measured curve, again consistent with a visual inspection of Fig.4e. It is also noted that the overall TAT is within 2.2% of the measured value for the best ERD/TAT and within 14.4% for the worst TAT.

Table 2. Metric values for maritimeEXODUS prediction of SGVDS2

		SC					n	ERD	EPC	% diff TAT
	s/n	0.01	0.02	0.03	0.04	0.05				
BEST TAT	**Overall**	**0.9**	**1.0**	**1.0**	**1.0**	**1.0**	**1743**	**0.11**	**1.1**	**-0.1**
	AS A	**0.6**	**0.9**	**0.9**	**0.9**	**0.9**	**397**	**0.14**	**1.1**	**-0.1**
	AS B	**0.9**	**0.9**	**1.0**	**1.0**	**1.0**	**561**	**0.11**	**1.0**	**2.2**
	AS C	**0.7**	**0.8**	**0.8**	**0.9**	**0.9**	**434**	**0.12**	**1.1**	**11.8**
	AS D	**0.5**	**0.8**	**0.9**	**0.9**	**0.9**	**351**	**0.18**	**1.1**	**4.1**
Worst TAT	**Overall**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1743**	**0.12**	**1.1**	**-14.4**
	AS A	**0.7**	**0.9**	**0.9**	**0.9**	**0.9**	**397**	**0.16**	**1.2**	**-16.2**
	AS B	**0.9**	**1.0**	**1.0**	**1.0**	**1.0**	**561**	**0.11**	**1.0**	**-8.3**
	AS C	**0.8**	**0.9**	**0.9**	**1.0**	**1.0**	**434**	**0.11**	**1.1**	**-0.3**
	AS D	**0.6**	**0.8**	**0.9**	**0.9**	**0.9**	**351**	**0.18**	**1.1**	**-5.0**
Best ERD	**Overall**	**0.9**	**1.0**	**1.0**	**1.0**	**1.0**	**1743**	**0.08**	**1.1**	**-2.2**
	AS A	**0.8**	**0.9**	**0.9**	**0.9**	**0.9**	**397**	**0.13**	**1.1**	**-18.0**
	AS B	**0.8**	**0.9**	**0.9**	**0.9**	**1.0**	**561**	**0.10**	**1.0**	**-5.7**
	AS C	**0.8**	**0.8**	**0.9**	**0.9**	**0.9**	**434**	**0.10**	**1.1**	**9.5**
	AS D	**0.8**	**0.9**	**0.9**	**0.9**	**0.9**	**351**	**0.15**	**1.0**	**8.7**

Next consider the shape of the predicted AS arrival curves. For each of the three cases, the predicted assembly curves for each AS show very good agreement with the experimental data. For an s/n of 0.02, the SC values for each AS are close to 1.0 (SC $\geq$ 0.8). This suggests that the shapes of the predicted assembly curves are in good agreement with the measured curves, again supporting the conclusions of the visual inspection. This s/n value, representing 2% of the data-set, is larger than that for the overall assembly curve, but is still considered small. For the smallest of the AS data-sets (AS D), this represents an s value of 7, while for the largest of the AS data-sets (AS B), this represents an s value of 11. These observations are consistent with a visual inspection of Fig.4. which suggests that the shapes of all the predicted AS arrival curves are in good agreement with the shape of the measured curves.

Next consider the magnitude of the difference between the predicted and measured AS arrival curves. The ERD values for each AS are quite low (< 0.20), with that for AS D in all cases being the greatest (0.18). Finally, each of the three cases produce good values of EPC, with all values being close to 1.0. The worst value is for AS A for the worst TAT case where EPC is 1.2. These values suggest that the predicted values of all the AS are reasonably close to the measured values, with the worst TAT case producing the poorest results. These observations are again consistent with a visual inspection of Fig.4. Indeed, the metric values suggest that AS B produces the best overall agreement with the measured values which is arguably supported by a visual inspection of the curves in Fig.4.

Based on this analysis, a set of acceptance criteria can be defined for SGVDS2 that takes into consideration the uncertainties in the experimental data and confirms that the maritimeEXODUS predictions presented in Fig.4 are arguably a good match for the experimental data based on a visual inspection. A general two-step validation protocol is suggested based in part on the philosophy of MSC 1238, which currently only focuses on the overall assembly time. In the first step of the validation protocol, the acceptance criteria are applied to the model predictions of the overall assembly. To be deemed acceptable, the model predictions must satisfy all elements of the acceptance criteria. If successful, the second step of the validation protocol is considered. In the second step, the acceptance criteria are applied to each of the four AS with a

minimum of 10 passes out of a possible 12 being deemed to be acceptable. Furthermore, no more than one failure can occur in any one AS. The validation protocol and acceptance criteria are applied to the model predictions which produce the best ERD. If the protocol is applied in this manner and the software meets the criteria, it demonstrates that the software is capable of producing an acceptable level of agreement with the experimental data for the entire assembly process. The suggested acceptance criteria are as follows:

(i) ERD ≤ 0.25
(ii) $0.8 \leq$ EPC ≤ 1.2
(iii) SC ≥ 0.8 with s/n = 0.02
(iv) Predicted TAT for the overall assembly to be within 15% of the measured value. This criterion is only applied to step 1 of the acceptance process.

Applying the suggested validation protocol to the maritimeEXODUS data presented in Table 2, we note that in the first step the model predictions satisfy all four criteria and hence the second step of the validation protocol is considered. In the second step each AS satisfies all the criteria. As the model predictions have satisfied all four criteria in step 1 and 12 of the 12 criteria in step 2, the model is considered to have satisfied the acceptance criteria.

DISCUSSION

The results presented in this paper are for blind predictions of the evacuation performance of the CS. By necessity, when used by other researchers, the comparisons will not be blind as the results will have been published. However, this is not considered to detract from the value of the validation data-sets. Indeed as the geometry, starting locations of the population, population response times and population end points are specified as part of the validation data-set, and all other model parameters are specified by MSC1238, there is little opportunity to tune the evacuation model to produce ideal results. However, due to the nature of the data in the validation data-set, it is possible for users to continually run batches of 50 simulations until an appropriate best ERD case is produced i.e. one that satisfies the criteria. This is due to not knowing the exact starting location of each agent and because the precise response time for each agent is not known, thus each simulation randomly produces a different allocation of response times and precise starting locations, some of which may be more favourable than others. To explore this possibility two additional batches of 50 simulations were produced for SGVDS2 using maritimeEXODUS and the results from the metric analysis are presented in Table 3 and 4.

Table 3. Metric values for maritimeEXODUS prediction of SGVDS2 – batch 2 best ERD

	SC					n	ERD	EPC	% diff TAT
s/n	0.01	0.02	0.03	0.04	0.05				
Overall	**0.8**	**0.8**	**0.9**	**1.0**	**1.0**	**1743**	**0.09**	**1.0**	**-7.9**
AS A	**0.4**	**0.4**	**0.5**	**0.5**	**0.6**	**397**	**0.16**	**1.0**	**-6.6**
AS B	**0.9**	**0.9**	**0.9**	**0.9**	**1.0**	**561**	**0.10**	**1.0**	**-12.0**
AS C	**0.7**	**0.9**	**0.9**	**1.0**	**1.0**	**434**	**0.11**	**1.0**	**16.9**
AS D	**0.7**	**0.8**	**0.9**	**0.9**	**1.0**	**351**	**0.13**	**1.0**	**-1.2**

From the results presented in Table 3 and 4, the results for the SC for batch 3 are marginally better than for batch 1 (see Table 2) while the results for batch 2 are marginally worse than batch 1. All the SC values for batch 3 satisfy the acceptance criteria, while in batch 2, the SC for AS A fails the criteria. The ERD values for batch 3 and 2 are marginally worse than for batch 1, with all the ERD values satisfying the acceptance criteria. The EPC values for batch 2 are marginally better than those for batch 1, while batch 3 are similar to those for batch 1. The largest variation in parameters between the three batches of results occurs for the time for the last agent to assemble overall and in each AS i.e. TAT. In batch 1, the overall TAT is under-predicted by 2.2%, while in batch 2 it is under-predicted by 7.9% and in batch 3 it is under-predicted by 7.2%. The greatest difference in the AS TAT occurs for AS D, where batch 1 over-predicts by 8.7% while batch 2 under-predicts the TAT by 1.2% and batch 3 under-predicts by 2.3% - a maximum difference of some 11%. However, as this criteria is only applied to the overall assembly results, all three cases are considered acceptable. Nevertheless, the large variation in the TAT for the AS demonstrates that the TAT is not a reliable measure. Due to the random allocation of precise starting location and response times, it is

possible that an agent is assigned a starting location which results in the furthest possible travel distance and the longest possible response time creating an abnormally long TAT. Furthermore, should that agent be associated with the AS that takes longest to assemble; it could severely impact the overall TAT. This is why the percentage difference in the TAT criteria should not be applied to individual AS, and if it is used at all, it should only be applied to the overall TAT.

Table 4. Metric values for maritimeEXODUS prediction of SGVDS2 – batch 3 best ERD

	SC					n	ERD	EPC	% diff TAT
s/n	0.01	0.02	0.03	0.04	0.05				
Overall	1.0	1.0	1.0	1.0	1.0	1743	0.09	1.1	-7.2
AS A	0.8	0.9	0.9	0.9	0.9	397	0.14	1.1	-10.7
AS B	0.9	1.0	1.0	1.0	1.0	561	0.09	1.0	-1.8
AS C	0.8	0.8	0.9	1.0	1.0	434	0.10	1.0	19.3
AS D	0.8	0.9	0.9	0.9	0.9	351	0.15	1.1	-2.3

Based on the metric values, while there are some differences in the precise values for the three components of the metric, the same conclusion with respect to acceptability would be made. Arguably, the results for batch 3 are marginally the best, while the results for batch 2 are marginally the worst. However, these observations cannot be generalised to other software tools and so there is some room for users to optimise their results.

While the best level of agreement between the numerical predictions and the experimental results for the overall assembly curve is identified by selecting the curve with the best ERD, the worst level of agreement can also be identified by selecting the curve with the worst ERD. If the validation protocol is applied in the worst ERD case (i.e. case producing the largest ERD value) and the software meets the acceptance criteria, it demonstrates that the worst of the 50 simulations are deemed to be satisfactory and so it is reasonable to expect that all 50 simulations within the batch – at least concerning the prediction of the overall assembly - will be acceptable. Presented in Table 5 are the metric values of the worst ERD case for both validation data-sets. As can be seen both sets meet the acceptance criteria associated with each validation data-set.

Table 5. Metric values for maritimeEXODUS prediction of worst ERD for SGVDS1 and SGVDS2

		SC					n	ERD	EPC	% diff TAT
	s/n	0.01	0.03	0.05	0.07	0.09				
SGVDS1 Worst ERD	Overall	0.8	0.9	0.9	1.0	1.0	480	0.34	1.1	-29.4
	AS A	0.3	0.5	0.8	0.8	0.8	77	0.39	1.3	-42.4
	AS B	0.4	0.7	0.7	0.8	0.9	142	0.37	1.1	-29.2
	AS C	0.2	0.4	0.6	0.7	0.8	74	0.29	1.3	-28.4
	AS D	0.7	0.8	0.9	0.9	0.9	187	0.57	0.7	-22.3
SGVDS2 Worst ERD	s/n	0.01	0.02	0.03	0.04	0.05	n	ERD	EPC	% diff TAT
	Overall	1.0	1.0	1.0	1.0	1.0	1743	0.15	1.1	-11.4
	AS A	0.7	0.8	0.9	0.9	0.9	397	0.19	1.2	-15.6
	AS B	0.9	0.9	1.0	1.0	1.0	561	0.12	1.1	-16.3
	AS C	0.6	0.7	0.8	0.9	0.9	434	0.15	1.1	13.9
	AS D	0.6	0.8	0.9	0.9	0.9	351	0.21	1.2	-5.8

Finally, the validation protocol described above is being applied blind to two other commonly used ship evacuation models, EVI [16] and ODIGO [17] (by their developers), as part of the SAFEGUARD project. Once this analysis is completed the validation protocol and acceptance criteria for SGVDS1 and SGVDS2 will be finalised. A more complete description of the validation data set and the suggested validation protocol can be found at: http://bit.ly/1eGeYEa.

CONCLUSIONS

Data from a semi-unannounced assembly trial at sea for a cruise ship have been collected consisting of passenger; response time data, starting locations, end locations and arrival time at the designated assembly stations. The response time data was collected using digital video cameras while the start and end locations and the arrival time for the passengers was collected using a novel Infra-Red (IR) data acquisition system consisting of ship-mounted IR beacons and IR data logging tags worn by each passenger. The collected data is used to define two unique validation data-sets for ship evacuation models. The data-sets are considered unique for a number of reasons, primarily because unlike most validation data-sets, they contain information defining; occupant response times, starting locations, end locations and final arrival times. Furthermore, the trials were conducted on real ships, at sea and were semi-unannounced making the results relevant, credible and realistic.

A validation protocol and acceptance criteria have been proposed based on the collected data. The acceptance criteria are objective and are determined by a metric consisting of three measures, the Euclidean Relative Difference, Euclidean Projection Coefficient and Secant Cosine. Collectively the metric measures the magnitude of the distance between the predicted and experimental data and the similarity of the shapes of the predicted and experimental arrival time curves. The proposed acceptance criteria take into consideration uncertainties associated with the measured data in each of the data-sets.

In blind applications of the validation protocol to the maritimeEXODUS ship evacuation software, the software was found to satisfy the acceptance criteria for the data-set, suggesting that it is capable of predicting the outcome of the assembly process for these two vessels to the specified level of accuracy as defined by the acceptance criteria. This work is being continued with the application of the validation protocol to two other evacuation tools, EVI and ODIGO.

It is proposed that the suggested validation protocol and the acceptance criteria could be used by IMO as part of a validation suite to determine acceptability of maritime evacuation models in a future enhancement to MSC1238. In this way we hope to improve the reliability of the assessment of ship evacuation capabilities based on computer simulation and hence the safety of all those who travel and work on passenger ships.

ACKNOWLEDGEMENT

The SAFEGUARD project (contract 218493) is funded under the European Union Framework 7 Transport initiative. The authors acknowledge the co-operation of their project partners.

REFERENCES

[1] IMO, “Interim Guidelines for Evacuation Analyses for New and Existing Passenger Ships”, IMO MSC/Circ 1033, 6 June 2002.

[2] “Guidelines for Evacuation Analysis for New and Existing Passenger Ships”, IMO MSC/Circ 1238, 30 Oct 2007

[3] IMO Fire Protection Sub-Committee, 51st session, Work Package 3, FP 51/WP.3, 8 Feb 2007.

[4] Brown, R., Galea, E.R., Deere, S., and Filippidis, L., “Passenger Response Time Data-Sets for Large Passenger Ferries and Cruise Ships Derived from Sea Trials”, The Transactions of the Royal Institution of Naval Architects, International Journal of Maritime Engineering, ISSN 1470-8751, Vol 155, Part A1, pp33-47, 2013.

[5] Galea, E.R., Deere, S., Brown, R., and Filippidis, L., “An Evacuation Validation Data Set for Large Passenger Ships”, To appear, Pedestrian and Evacuation Dynamics 2012. 6th International Conference. Proceedings. June 6-8, 2012, Springer, New York, NY. 2013

[6] Galea, E.R., Brown, R.C., Filippidis. L., and Deere. S.: Collection of Evacuation Data for Large Passenger Vessels at Sea, Pedestrian and Evacuation Dynamics 2010. 5th International Conference. Proceedings. March 8-10, 2010, Springer, New York, NY, Peacock, R.D., Kuligowski, E.D., and Averill, J.D., Editor(s), 163-172, (2011).

[7] Deere, S., Galea, E.R., Lawrence, P., Filippidis, L. and Gwynne, S.: The impact of the passenger response time distribution on ship evacuation performance, International J of Maritime Eng, Vol 148, Part A1, 35-44, (2006).

[8] Galea, E.R., Lawrence, P., Gwynne, S., Sharp, G., Hurst, N., Wang, Z., and Ewer., J., "Integrated fire and evacuation in maritime environments", Proc of the 2nd Int Maritime Safety Conference on Design for Safety, Sakai Japan, Publisher Ship and Ocean Foundation, 27-30 Oct 2004, pp 161-170.

[9] Boxall, P., Gwynne, S., Filippidis, L., Galea, E.R. and Cooney. D., "Advanced Evacuation Simulation Software and its use in Warships", Proc of the Human Factors in Ship Design, Safety and Operation, London UK, Publisher The Royal Institute of Naval Architects, 23-24 Feb 2005, pp 49-56.

[10] Caldeira-Saraiva, F., Gyngell, J., Wheeler, R., Galea, E.R., Carran, A., Skjong, R., Vanem, E., Johansson, K., Rutherford, B., and Simoes, A.J., "Simulation of ship evacuation and passenger circulation", Proc 2nd Int Maritime Safety Conference on Design for Safety, Sakai Japan, Publisher Ship and Ocean Foundation, 27-30 Oct 2004, pp 197-205.

[11] Deere, S J, Galea, E R and Lawrence, P, "A Systematic Methodology to Assess the Impact of Human Factors in Ship Design", Applied Mathematical Modelling, Applied Mathematical Modelling, 33, 867-883, 2009. <http://dx.doi.org/10.1016/j.apm.2007.12.014>

[12] Andrews, D J, Pawling, R, Casarosa, L, Galea, E R, Deere, S and Lawrence, P, "Integrating Personnel Movement Simulation into Preliminary Ship Design", International Journal of Maritime Engineering, Volume 150 Part A1 pp 19-34, ISSN 1479-8751, 2008. http://www.rina.org.uk/ijme0801.html

[13] Galea, E.R., Deere, S., Sharp, G., Filippidis, L., Lawrence, P., and Gwynne, S.: Recommendations on the nature of the passenger response time distribution to be used in the MSC 1033 assembly time analysis based on data derived from sea trials." International J of Maritime Eng, Vol 149, Part A1, 15-29, (2007).

[14] Peacock, R.D., Reneke, P.A., Davis, W.D., Jones, W.W.: Quantifying Fire Model Evaluation Using Functional Analysis, Fire Safety Journal, 22, 167-184, (1999).

[15] Peacock, R.D.: Private Communication with E.R.Galea, 23 November 2011.

[16] Vassalos, D., Kim, H., Christiansen, G., Majumder, J., 'A Mesoscopic Model for Passenger Evacuation in a Virtual Ship-Sea Environment and Performance-Based Evaluation', Pedestrian and Evacuation Dynamics – April 4-6, 2001 – Duisburg. pp369-391. ISBN: 3-540-42690-6, (2001).

[17] Pradillon, J.Y., 'ODIGO - Modelling and Simulating Crowd Movement onboard Ships', 3rd International Conference on Computer and IT Applications in the Maritime Industries, COMPIT, Siguenza, Spain, pp278-289, 2004.