The Use of Confidence Intervals to Determine Convergence of the Total Evacuation Time for Stochastic Evacuation Models

ANGUS GRANDISON*, STEVEN DEERE, PETER LAWRENCE and EDWIN RICHARD GALEA

Fire Safety Engineering Group (FSEG)

University of Greenwich

Old Royal Naval College

Park Row

Greenwich

LONDON SE10 9LS, UK

*Corresponding Author: A.J.Grandison@gre.ac.uk

ABSTRACT

International guidelines (IMO MSC.Circ 1533) specify that evacuation models used to certify evacuation performance of passenger ships must demonstrate that the calculated representative evacuation time, the *sample* 95th percentile time τ^S, is lower than a prescribed Pass/Fail Criterion Time (PFCT). In this paper a Confidence Interval Convergence Test (CICT) method is presented that minimises the computational burden required to demonstrate that a model design has passed/failed by calculating a CI for the population 95th percentile time, τ^P, rather than simply relying on τ^S determined from an arbitrary sample of 500 simulations as specified in the current guidelines. The CICT has comparable pass/fail accuracy to using 500 simulations whilst significantly reducing the number of simulations required when the PFCT is far from the τ^P. In addition, the proposed method has superior accuracy to the convergent method described in the IMO guidelines. Furthermore, the methodology described in the guidelines fails to identify situations where there may be uncertainty in the pass/fail status due to proximity of τ^P to PFCT. The CICT identifies these situations and provides a means of resolving the uncertainty. The CICT can be applied to any stochastic evacuation model to determine parameter convergence.

KEYWORDS: 95th percentile, confidence intervals, evacuation modeling, convergence, maritime safety, IMO.

Nomenclature		n	sample size of simulations
CI	Confidence Interval	R_{CI}	range of the CI (s)
CI(x%)	CI with an x% confidence level	T_b	TET of simulation b (s)
CICT	Confidence Interval Convergence Test	τ	95th Percentile TET (s)
IMO	International Maritime Organisation	τ^P	*population* 95th Percentile TET (s)
PFCT	Pass/Fail Criterion Time	$\tau^{S(=n)}$	*sample* (of size *n*) 95th Percentile TET (s)
TET	Total Evacuation Time (s)		

1 INTRODUCTION

Many evacuation simulation models (Gwynne et al., 1999; Kuligowski et al. 2010) employ a stochastic approach for the representation of behaviour and movement (Gwynne et al., 2001, 2003; Ha et al., 2012; Korhonen et al., 2008; Meyer-König, 2005; Park et al., 2004; Pradillon, 2003; Thompson and Marchant, 1995; Vassalos et al., 2002) as they attempt to reflect the probabilistic nature of human behaviour (Averill, 2011). This is consistent with real behaviour since if any evacuation experiment is repeated using the same population and same starting conditions it is likely that the evacuation will progress differently and result in a different total evacuation time (TET). However, two key questions that arise when using a stochastic evacuation model concerns how many simulations are required to obtain a given level of confidence that the predicted results provide a true indication of the expected outcome for the scenario and what should be considered the representative value of predicted parameters such as TET for a given scenario. Given a distribution of predicted TETs there are a number of possible candidate values for the representative TET such as the longest TET, the mean TET, the median TET, or the 95th percentile TET. To a certain extent, the predicted parameter used to represent the distribution of possible results is dependent on the purpose for undertaking the analysis. If it is part of a risk analysis, it may be appropriate to take a reasonable worse case and so the 95th percentile TET may be appropriate, if the analysis is more concerned with typical performance, then the mean TET may be appropriate.

While there has been some interest in these issues (Meacham et al., 2004; Ronchi et al., 2014) for building applications, there are currently no internationally agreed guidelines on how to address this issue for building applications. However, the International Maritime Organization (IMO) in their guidelines for evacuation analysis (IMO, 2016) specifies that when assessing the evacuation capability of a passenger ship using an advanced egress model, a minimum of 500 simulations must be performed and that the representative TET is the 95th percentile TET, τ, from those simulations. The possible use of the 95th percentile (of a sample of simulations) TET has also been suggested for the building (Meacham et al., 2004) and aviation (Galea, 2006; Galea et al., 2010) industries. The IMO (2016) guidelines further stipulate the minimum number (four) and nature of scenarios that must be investigated for each new ship design. This includes the nature of the population (of agents) distribution (age, gender and number of disabled occupants) and the range and distribution of key parameters such as occupant response times and walking speeds. In addition the guidelines stipulate that each scenario must be repeated with the key parameters varied between the given ranges for each repeat simulation. Thus, in addition to the natural variation in evacuation output that can be expected due to the stochastic nature of behaviour (even if none of the input parameters are altered), varying the key parameters between each of the repeat simulations will result in even greater variation in the predicted output. As stochastic evacuation models generally use pseudo-random numbers then there will be a finite number of possible different simulations that can be produced, but the number of unique simulations could be very large, $2^{19937}–1$ ($>10^{6001}$) if a Mersenne-Twister Random Number Generator (RNG) (Matsumoto and Nishimura, 1998) or 2^{249} ($>10^{74}$) for a R250 (Kirkpatrick and Stoll, 1981) RNG is used. From a practical point of view it is generally only possible to take a relatively small sample (<10,000) of all possible simulations and so the population of simulations that these are drawn from is effectively infinite.

Prior to the recently updated IMO guidelines, IMO (2007) specified that the 95th percentile value from a 50 simulation trial sample, $\tau^{S=50}$, was sufficient to represent the predicted evacuation time for the vessel design. When undertaking an evacuation analysis, the representative TET is compared to the relevant **Pass Fail Criterion Time** (PFCT) and the design is deemed to have passed if the τ is less than the PFCT. However, the variability of τ between samples was not examined and there is no requirement for error bars to be specified for the representative value. Thus, in the previous IMO guidelines, $\tau^{S=50}$ was assumed to be a good estimation of the 95th percentile value of the entire population of predicted TETs for the given scenario, τ^{P}. However, there is considerable variation in $\tau^{S=50}$ (see Fig. 1) and using $\tau^{S=50}$ to represent τ^{P} can lead to an increasing number of false positives (type I error where a poor design is deemed to have passed) and false negatives (type II error where a satisfactory design is deemed to have failed) as τ^{P} for the vessel design and scenarios gets closer to the pass fail criterion time, PFCT. It is noted that the actual τ^{P} is generally impractical to determine as it would require running a very large number of simulations to ensure that all possible permutations of model input parameters and all the natural inherent model variability was accounted for. It is further noted that the τ^{P} for a particular model cannot be assumed to exactly represent reality due to assumptions used to specify the artificial benchmark scenarios, the simplifications within the model and a lack of data defining the performance of the population in general and particularly in emergency situations. The IMO (2016) guidelines add a 25% safety factor to account for these uncertainties.

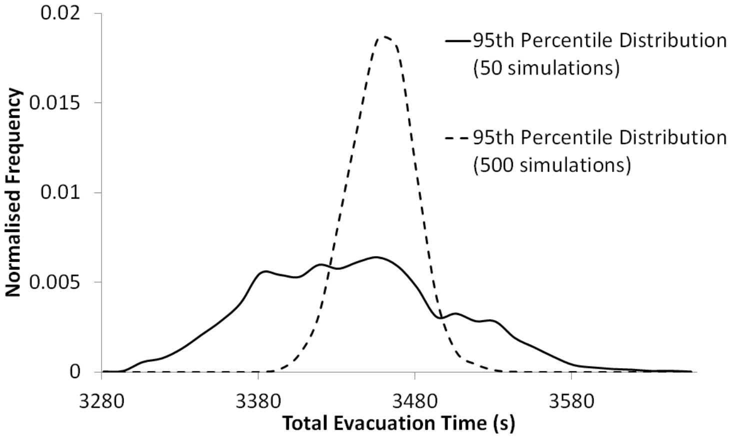


Fig. 1. Example variability of τ^S across one million experiments for 50 and 500 simulation sized samples.

A false positive occurs when τ^S is less than the PFCT but τ^P is greater than the PFCT. Similarly, a false negative occurs when τ^S is greater than the PFCT but τ^P is less than the PFCT. In the most recent version of the guidelines, IMO (2016) attempted to address this problem by increasing the sample size to a minimum of 500 simulations. In this approach it is assumed that $\tau^{S=500}$ is likely to be a more precise estimate of τ^P thereby reducing, but not eliminating, the probability of false positives or false negatives. This can be seen in Fig. 1, where the difference between the maximum and minimum values of $\tau^{S=50}$ for a sample involving 1 million simulation experiments is 540s, compared to the more precise extent of $\tau^{S=500}$ which is 191s.

However, while the IMO (2016) Guidelines state that the 500 simulations are considered a minimum, they provide no advice as to what circumstances may require additional simulations to be considered. The inevitable effect of this omission is that most engineers will treat the stated minimum as effectively the required number of simulations. They may be motivated to undertake more simulations in the event that the design failed. Furthermore, performing 500 simulations is potentially a considerable computational burden when evaluating the design of a large passenger ship with many thousands of passengers and so engineers are unlikely to voluntarily perform more simulation unless required to. This is considered a serious omission as no proof is required to demonstrate that the sample $\tau^{S=500}$ provides a good representation of the population τ^P.

Given that performing 500 simulations may be more than required in some cases and acknowledging the computational burden of undertaking the task, IMO provided the option of performing fewer than the specified 500 simulations if it could be demonstrated that the sample 95th percentile time had converged, as stated in the IMO (2016) guidelines, "*The minimum of 500 different simulations can be reduced when a convergence is determined by an appropriate method...*". Within the guidelines a suggested convergent method that increases the precision that is required for τ^S as the PFCT gets closer to τ^S is presented however, the efficiency of this approach is not discussed.

While the IMO Guidelines provide a means for demonstrating that fewer than 500 simulations may be required, it does not provide a means for demonstrating that 500 simulations may be insufficient. Thus a motivation for this paper is to provide a methodology that has comparable pass/fail accuracy as using 500 simulations whilst minimising the computational burden required when a design clearly passes or fails the PFCT by a significant margin and which indicates that more than 500 simulations may be required to make a decision on the suitability of the design.

Ronchi et al. (2014) have proposed convergence criteria for stochastic evacuation models based on five measures. The first two measures are based on comparing the difference between the mean and standard deviation of TETs for *j* simulations against the mean and standard deviation obtained for *j*-1 simulations against a specified tolerance. The other three measures are based on functional analysis (Peacock et al., 1999) which compare properties of the average overall egress curve (i.e. the number of exited agents vs time) for *j* simulations against properties of the average overall egress curve for *j*-1 simulations (note that the metrics specified in Peacock et al. (1999) are incorrectly specified and are corrected in Galea et al. (2013)). In their work the representative TET is the mean value of all the TETs generated together with the standard deviation of the TETs and is therefore, in its current form, unsuitable for examining the convergence of τ. Galea et al. (2012, 2013, 2014) have also used functional analysis for validating a computed egress curve against an experimentally derived egress curve for a large cruise ship. The work presented in this paper differs from these by considering convergence based on confidence intervals (CIs) and particularly the 95th percentile TET, τ. Ronchi et al. (2014) note that their concepts of convergence are a potential limitation of their method. They argue that this limitation is tempered by the simplicity of the method. In this paper we argue that comparing the difference between successive statistical measures as utilised in Ronchi et al (2014), while appropriate for iterative numerical solvers for deterministic variables are not ideal for stochastic simulations making them less reliable than the CI methods proposed in this paper, which are as simple to apply as the alternative methods and give a valid statistical interpretation.

Finally, it is important to note that convergence for one parameter - whether it is the mean TET, median TET, τ or some other population parameter - does not guarantee convergence for other predicted parameters of interest, such as measures of congestion or casualty numbers to the same accuracy or precision. Therefore an engineer needs to ensure that the required levels of convergence are achieved for all parameters that may be of interest to them. For example, the mean TET is likely to converge with fewer simulations than τ for an equal level of precision. The methods described in this paper only refer to single parameter convergence but this does not preclude its use as part of a multi-objective procedure if an engineer wishes to pursue this approach.

2 METHODOLOGY

Within the IMO (2016) guidelines, there is currently no stated requirement to have error bars or CI associated with the predicted representative evacuation time. CIs are conventionally simple to calculate when the population parameter is the mean of the values by using standard statistics (e.g. the Central Limit Theorem (Rice, 1995) but is less obviously derived for percentile style parameters. This is the case for stochastic evacuation simulations where the underlying population distribution of TETs is unknown. If a CI can be established then it is possible to reduce the number of simulations required to identify whether or not a particular design's τ^P is actually greater than or less than the PFCT.

2.1 Confidence Intervals

In statistics, a **CI** (Neyman, 1937) is a type of interval estimate of a population parameter (i.e. τ^P). It is an observed interval (i.e. it is calculated from the observations), that generally differs between samples, that frequently includes the value of an unobservable parameter of interest if the experiment is repeated. The frequency that the observed interval contains the parameter is determined by the **confidence level**. If CIs are constructed across many separate data analyses of replicated experiments, the proportion of such intervals that contain the true value of the parameter will match the given confidence level. This is illustrated for a 95% confidence level in Fig. 2 where the vast majority, 47 (~95%), of the sample CIs contain τ^P but 3 (~5%) of the sample CIs do not. It is impossible to guarantee that a sample contains τ^P let alone a particular CI for that sample but a confidence level of *x*% equates to *x*% of the sample CIs containing τ^P.

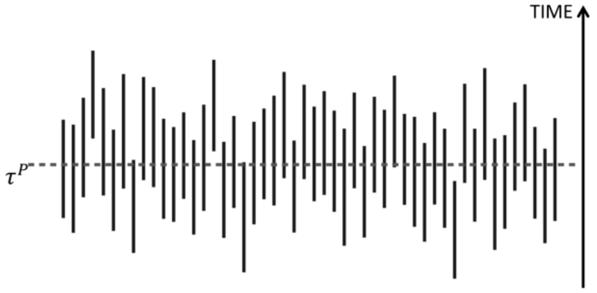


Fig. 2. 95% Confidence Intervals for τ^P for 50 different samples.

When the underlying population distribution of TETs and τ^P is unknown, the CI can be established using either a Bootstrap (resampling) method (Efron and Tibishirani, 1993) or by recasting the sample into a set of Bernoulli trials (Papoulis, 1984) (a trial with two possible outcomes). Both methods were tested and found to perform similarly but recasting into Bernoulli trials leads to a simpler method for the end-user and is presented in this paper.

The parameter of interest is τ^P and although its actual value is generally unknown the probability of a single trial having a TET less than or greater than τ^P is known by definition. The probabilities are 0.95 and 0.05 respectively. This is now a Bernoulli trial that can be treated using a binomial distribution in much the same way as computing the probability for a sequence of coin tosses. This binomial distribution can then be used to produce a confidence interval for τ^P for a sample of simulation TETs.

The probability mass function P that exactly k simulations are less than τ^P in n trial simulations is given by Eq.1.

$$P(n,k) = \binom{n}{k} 0.95^k 0.05^{n-k} \tag{1}$$

Where,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{2}$$

This is plotted in Fig. 3, for a total sample size of 250 simulations.

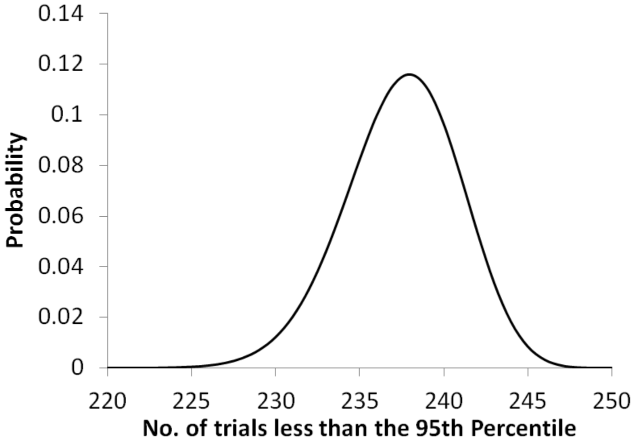


Fig. 3. Probability Mass Function for the number of TETs less than τ^P for 250 simulations.

By visual inspection of Fig. 3 it can be seen that the probability that the 250 simulation TET sample has 225 simulation TETs or fewer below τ^P is very small and the probability that the sample has 247 simulation TETs or more below τ^P is also very small. For any particular sample of 250 simulations it is highly likely that there will be between 225 and 248 simulations less than τ^P. This interval is more formally described as a CI.

2.1.1 Constructing Confidence Intervals

The CI is chosen so that the cross-hatched regions in Fig. 4 are equal in area to each other. For a confidence level of 99% (0.99) the area outside of the interval (α) is 1% (0.01) and therefore 0.5% (0.005) either side. The lower bound of the CI is calculated by finding the value of *i* that satisfies (Eq. 3):-

$$\sum_{k=0}^{i} \binom{n}{k} 0.95^k 0.05^{n-k} = \alpha/2 \tag{3}$$

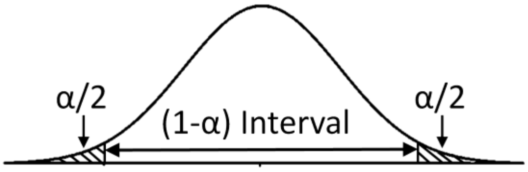


Fig. 4. Example confidence interval.

Similarly the upper bound of the CI is calculated by finding the value of *j* that satisfies (Eq. 4):-

$$\sum_{k=j}^{n} \binom{n}{k} 0.95^k 0.05^{n-k} = \alpha/2 \tag{4}$$

For the discrete binomial distribution it is generally not possible to find an exact match so the lower bound is taken as the largest value of *i* that gives a summation S that does not exceed $\alpha/2$. Similarly for the upper bound the lowest value of *j* that gives a summation that does not exceed $\alpha/2$. This leads to a greater confidence level than the stated $(1 - \alpha)$ (Clopper and Pearson, 1934). The CI could have been obtained by using an approximation such as the Normal, Wilson (1927) or Agresti and Coull (1998) intervals but these intervals may be inaccurate or have a coverage that is less than the stated confidence level (Brown et al., 2001).

2.1.2 Order Statistics

Now the predicted TETs (shortened to *T* in mathematical expressions and equations) from the trials are ordered $T_1 < T_2 < T_3 \ldots < T_{n-1} < T_n$. For the lower confidence limit where *i* trials are lower than τ^P then the lower confidence limit of τ^P must lie between T_i and T_{i+1}. The most conservative approach is to take the lower value as the estimate, i.e. T_i. For the upper confidence limit where *j* trials are lower than the τ^P then the upper confidence limit of τ^P must lie between T_j and T_{j+1}. The most conservative approach is to take the upper value as the estimate, i.e. T_{j+1}. Similarly, τ^S lies somewhere between the two consecutive values $T_{\lfloor 0.95*(sample\ size)\rfloor}$ and $T_{\lfloor 0.95*(sample\ size)\rfloor+1}$, where $\lfloor x \rfloor$ returns the integer truncation of x. Once again, the most conservative approach is to take the upper value, i.e. $T_{\lfloor 0.95*(sample\ size)\rfloor+1}$.

Table 1 gives the appropriate values to use for the estimate of the population 95th percentile, the lower and upper confidence limits for this value from the (ascending) ordered simulation times for a range of possible sample sizes. For a sample of 50 simulations it is impossible to define an upper confidence limit for the required confidence levels but the lower confidence limit can be defined; in fact there is a 7% probability that none of the simulations will be greater than τ^P. For a sample of 100 simulations it is also impossible to define an upper limit for a 99% or 99.9% confidence level but it can be defined for a 95% confidence interval. For 150 or more simulations it is possible to define both an upper and lower limit for all the specified confidence levels. For arbitrary sample sizes and CIs it is possible to calculate these values using a tool such as Microsoft EXCEL.

Table 1 - Ordered Simulation Times to use for the estimated 95th percentile and confidence limits

Number of Simulations	**Est of τ^P (=τ^S)**	**>95% Confidence Limit**		**>99% Confidence Limit**		**>99.9% Confidence Limit**	
		Lower Limit	**Upper Limit**	**Lower Limit**	**Upper Limit**	**Lower Limit**	**Upper Limit**
50	T_{48}	T_{43}	N/A	T_{42}	N/A	T_{40}	N/A
100	T_{96}	T_{89}	T_{100}	T_{88}	N/A	T_{86}	N/A
150	T_{143}	T_{136}	T_{148}	T_{134}	T_{149}	T_{131}	T_{150}
200	T_{191}	T_{183}	T_{197}	T_{180}	T_{198}	T_{178}	T_{199}
250	T_{238}	T_{229}	T_{245}	T_{227}	T_{246}	T_{224}	T_{248}
300	T_{286}	T_{277}	T_{293}	T_{274}	T_{295}	T_{270}	T_{297}
350	T_{333}	T_{323}	T_{341}	T_{320}	T_{343}	T_{317}	T_{345}
400	T_{381}	T_{370}	T_{389}	T_{367}	T_{391}	T_{363}	T_{394}
450	T_{428}	T_{417}	T_{438}	T_{414}	T_{439}	T_{410}	T_{442}
500	T_{476}	T_{464}	T_{485}	T_{461}	T_{488}	T_{457}	T_{490}
1000	T_{951}	T_{935}	T_{963}	T_{930}	T_{967}	T_{926}	T_{971}

2.2 Significance Testing

The procedure recently adopted by IMO (2016) is to determine whether the *sample* 95th percentile value of a 500 simulation sample ($\tau^{S=500}$) is lower (pass) or higher (fail) than the PFCT and is essentially the same procedure as used previously (IMO, 2007) except that in the earlier guidelines a minimum of 50 simulations was used. The alternative approach suggested in this paper, the CI convergence test (CICT), is similar to performing two one-tailed statistical significance tests and determines whether the PFCT is *significantly* lower (fail) or *significantly* higher (pass) than the *population* 95th percentile value (τ^P). The design will pass if the upper confidence limit TET is less than the PFCT and will fail if the lower confidence limit TET is greater than the PFCT. However, if the PFCT lies between the upper and lower confidence limits then it is undetermined whether the design has passed or failed as the τ^P also lies *somewhere* within the Confidence Interval and it is not possible to determine whether the PFCT is therefore greater than or less than the τ^P.

To illustrate the concept of the CICT consider a hypothetical example in which there are a sequence of values for PFCT (PFCT-1, PFCT-2, PFCT-3 and PFCF-4) which get progressively closer to τ^P as depicted in Fig. 5. A series of 50 CIs with a confidence level of 95%, that have been determined for the design, are also illustrated in Fig 5. In this hypothetical example, all the PFCT values are greater than τ^P for the underlying distribution so in principle the design under investigation should pass irrespective of which value of PFCT is used.

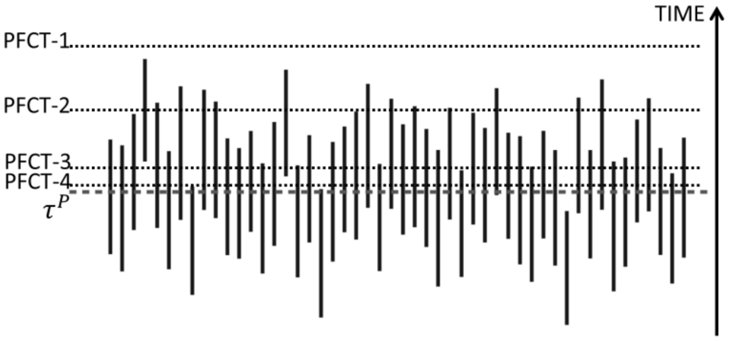


Fig. 5. 95% Confidence Intervals for 50 different samples plotted against τ^P and 4 PFCT values

Using the CICT outlined above, the samples are categorised as: 1) Pass, 2) Fail (False Negative), and 3) Undetermined for each of the PFCT levels (see Table 2).

Table 2 – Passed, Failed, Undetermined samples for different PFCT (Fig. 5) using CI

	Pass (CI Below PFCT)	**Fail (False Negative) (CI Above PFCT)**	**Undetermined (CI Through PFCT)**
PFCT-1	50	0	0
PFCT-2	36	0	14
PFCT-3	7	1	42
PFCT-4	3	2	45

It can be seen (Fig. 5 and Table 2) that all the samples will pass the PFCT-1 criterion but as the PFCT values gets progressively closer to τ^P fewer samples can be defined as a pass and more cases become undetermined and a small number of samples are defined as a fail (False Negative).

The proposed CICT can be further explored to determine what percentage of the samples would pass the upper confidence interval test. To simplify the analysis it is assumed that τ^S follows a normal distribution with a mean of τ^P and standard deviation ($0.5\times R_{CI(95\%)}/1.96$) where R_{CI} is the range of the CI (eq. 5). It is further assumed, for analytical convenience, that the upper bound of the 95% CI also follows a normal distribution with a mean of ($\tau^P + R_{CI(95\%)}/2$) and a standard deviation of ($0.5\times R_{CI(95\%)}/(1.96)$).

$$R_{CI} = T_{upper\ CI} - T_{lower\ CI} \tag{5}$$

Furthermore, to demonstrate this analysis we will make use of 10,000 simulated evacuation experiments of a cruise ship validation case (Galea et al., 2012, 2013, 2014), the details of which will be described later in this paper. From the set of 10,000 evacuation simulations, random samples of 1000 TETs were selected one million times. This is intended to represent one billion genuine repeat simulated evacuation experiments. Presented in Fig. 6 is the frequency distribution for the $\tau^{S=1000}$ and the upper and lower limits of the 95% confidence interval for the one million samples of 1000 repeat evacuation experiments. It can be seen, by visual inspection, that frequency distributions are approximately normal and so the fitted normal assumption for the distribution of τ^S and the upper 95% confidence limit is reasonable for this analysis.

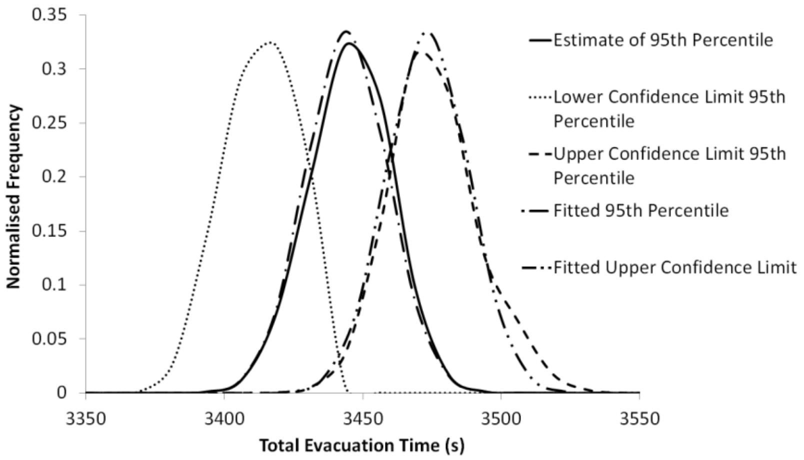


Fig. 6. Frequency Distribution for $\tau^{S=1000}$ and associated 95% Confidence Intervals and their fitted normal distribution approximations.

The normal fit will not necessarily be this close for all samples but the general behaviour explored in the following analysis should be similar. It is now possible to define the proportion of samples that would pass the test in terms of the ratio of (PFCT - τ^P) to $R_{CI(95\%)}$ by integrating this normal distribution defined in units of $R_{CI(95\%)}$ from -∞ to PFCT (see Fig. 7 and Table 3).

Table 3. Proportion of samples that pass for a given ratio of (PFCT - τ^P) to $R_{CI(95\%)}$ when comparing the PFCT to the upper CI (A: Upper CI limit < PFCT) and direct comparison with τ^S (B: τ^S < PFCT).

(PFCT –τ^P) ***$R_{CI(95\%)}$***	-1/5	-1/10	0	1/10	1/5	1/3	1/2	2/3	1	3/2
A	0.003	0.009	0.025	0.058	0.120	0.257	0.5	0.743	0.975	1.0
B	0.217	0.348	0.5	0.652	0.783	0.904	0.975	0.996	1.0	1.0

From Table 3 it can be seen that if PFCT = τ^P then 2.5% of all samples would be deemed to have passed; this is consistent with the definition of the 95% CI. The region where PFCT < τ^P represents the area where false positives can be obtained as the upper limit of the CI for a small proportion of samples will be less than τ^P. If PFCT < τ^P then the percentage of false positives tends to α/2 as (PFCT - τ^P)/R_{CI} tends to zero; the percentage of false positives tends to zero as (PFCT - τ^P)/R_{CI} tends to negative infinity. For a confidence level of 95% the maximum probability of false positives is less than 2.5%; for a confidence level of 99% the maximum probability of false positives is less than 0.5%; and for a confidence level of 99.9% the maximum probability of false positives is less than 0.05%. This compares very favourably to direct comparison of τ^S to the PFCT, as used in the former (IMO, 2007) and current guidelines (IMO, 2016), the probability of false positives would tend to 50% as |PFCT - τ^P| tends to zero. Samples that are false positives using the CICT would also have been false positives using direct comparison of τ^S to the PFCT but represent a small subset of all the possible false positives that would occur due to direct comparison of τ^S to the PFCT.

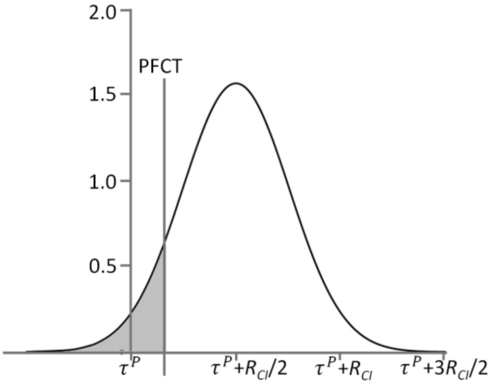


Fig. 7. Normal Distribution for upper limit of CI

When $R_{CI(95\%)}$ is much larger than |PFCT - τ^P| then only a small proportion of the samples will pass. When |PFCT - τ^P| is equal to $R_{CI(95\%)}/2$ then half the samples will pass. When |PFCT - τ^P| is equal to $R_{CI(95\%)}$ then 97.5% the samples will pass which is consistent with the 95% CI. When |PFCT - τ^P| is larger than $R_{CI(95\%)}$ then the vast majority of the samples will pass. An analysis for failing a sample can be obtained by applying a similar analysis to the lower bound of the CI.

As the PFCT gets closer to τ^P, for a fixed R_{CI}, the number of undetermined samples increases and the method needs to be extended to deal with these (Tables 2 and 3). From Table 3 it can be seen that more samples pass as R_{CI} becomes smaller relative to |PFCT - τ^P|. A feature of confidence intervals with a fixed confidence level is that R_{CI} tends to narrow (converge) with increasing sample size (see Fig. 8). For example the 95th Confidence Interval range for τ^P is the 89th and 100th percentile for a 100 simulation sample size compared to the 92.8th and 97th percentile for a 500 simulation sample size. More generally it can be shown that the R_{CI} in percentiles is inversely proportional to $\sqrt{n}$.

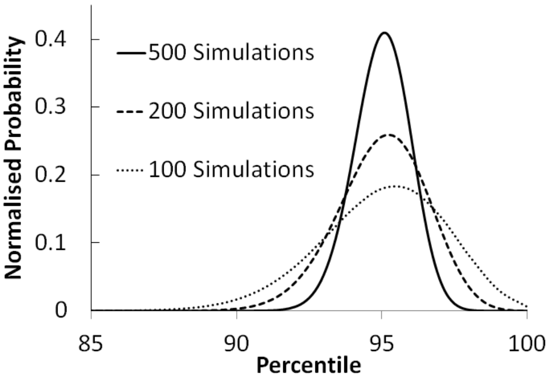


Fig. 8. Normalised Probability Density Functions for percentiles of TETs less than τ^P for 100, 200 and 500 total simulations in a sample.

This enables a convergent approach based on incremental testing where the size of the sample can be progressively increased until the sample's R_{CI} is narrow enough ($(R_{CI} \propto 1/\sqrt{n}) \rightarrow 0$ as $n \rightarrow \infty$) that it can be determined to have *significantly* failed or *significantly* passed. R_{CI} needs to be less than |PFCT - τ^P| to guarantee, to the specified confidence level, the pass or fail can be determined although this determination is likely to be achieved with a larger R_{CI} (see Table 3). It is always possible to resolve whether a sample has passed or failed, provided $\tau^P \neq$ PFCT, but as |PFCT - τ^P| tends to zero the number of simulations required to sufficiently narrow R_{CI} to identify whether the case passes (or fails) tends to infinity. Furthermore, the rate of convergence of R_{CI} is slow, e.g. halving R_{CI} requires the sample size n to be quadrupled. This is untenable and a maximum number of simulations (n_{max}) needs to be set with an additional criterion required to identify whether the case has passed or failed.

A series of tests were performed to obtain the average $R_{CI(95\%)}$ for a 'real' case, based on the cruise ship evacuation model validation test case (Galea et al., 2012, 2013, 2014), for a range of sample sizes. For each sample size one million experiments were performed and $R_{CI(95\%)}$ from these samples is averaged (see Table 4).

Table 4. Average $R_{CI(95\%)}$ for sample sizes 100 to 500

Sample Size	100	150	200	250	300	350	400	450	500
Average $R_{CI(95\%)}$	253.4s	162.6s	150.1s	132.8s	117.3s	106.9s	99.1s	92.9s	88.2s

From Table 4 it is observed that increasing the sample size reduces $R_{CI(95\%)}$ approximately in line with $1/\sqrt{n}$ although the $R_{CI(95\%)}$ is also dependent on the underlying TET distribution. A sample size of 500 simulations has an average $R_{CI(95\%)}$ of 88.2s meaning that if |PFCT - τ^P| is less than this then there is a significant probability that the size of sample required would reach the n_{max} limit (=500).

2.3 Suggested Methodology to Demonstrate Accurate Convergence

The proposed CICT methodology makes use of confidence intervals and ensures that the predicted total assembly time is determined accurately enough to determine whether it has passed or failed. Unlike the approach suggested within the IMO guidelines, the CICT not only provides a level of confidence to attach to the predicted assembly time, it can potentially achieve a similar level of accuracy in the predicted assembly time as the method in the IMO guidelines whilst utilising fewer simulations and hence computational time.

The CICT requires the generation of evacuation simulations in batches of 50 simulations. Each simulation is performed using its own randomly generated passenger populations to ensure there is no inadvertent biasing in the sample that could be caused by using the same initial passenger distribution for a number of simulations. Given the number of simulations performed, the representative times from the sample are selected using Table 1 and compared with the PFCT. If the test fails, another batch of 50 simulations is performed and combined with the results from the earlier set of simulations and retested. This process is continued until either the sample has passed or failed the test or the maximum number of simulations (n_{max}) has been reached. The process proceeds as follows:

1) Run a batch of 50 simulations, order the simulation times and test to see if the lower confidence bound e.g. T_{43} (of the τ^P estimate) is greater than the PFCT. If it is then the design has failed. Otherwise continue.

2) Run another batch of 50 simulations, order all 100 simulation times and test to see if the lower confidence bound e.g. T_{89} is greater than the PFCT. If it is then the design has failed. If not, test if the upper confidence bound e.g. T_{100} (95% CI only) is less than PFCT. If it is then the design has passed. Otherwise continue.

3) Run another batch of 50 simulations, order all 150 simulation times and test to see if the lower confidence bound e.g. T_{136} is greater than the PFCT. If it is then the design has failed. If not, test if the upper confidence bound e.g. T_{148} is less than the PFCT. If it is then the design has passed. Otherwise continue.

4) Repeat (3) adding another batch of 50 simulations for a total number of simulations of 200, 250, 300, 350, …, n_{max}-50, n_{max} until the design has passed or failed. n_{max} must be at least 500 (IMO, 2016) although it can be set to a higher value to give a greater level of pass/fail accuracy.

5) If design has not passed or failed after n_{max} simulations then the PFCT must be "close" and within the CI of τ^P. There are a number of ways this situation could be dealt with but it is suggested that the design passes if the $\tau^{S\geq500+}$ is less than the PFCT and fails otherwise. This is consistent with the suggested approach recently adopted by IMO (2016) when 500+ simulations are used.

It is recommended that the CI(95%) level is used (as this is specified within the current IMO (2016) guidelines) but when a case passes at this level it is also possible that it may also have passed a higher CI (i.e. 99% or 99.9%). In this case the result should be quoted to the highest passed CI level. The result should be quoted for the estimate of the 95th percentile, the CI (see Table 1), and the number of simulations performed. The use of a CI can be considered as an "error bar" on the estimate of the population 95th percentile. For example, results should be quoted as follows:

"Total Evacuation Time is 57m40s (i.e. T_{96}), 95%CI [55m11s, 60m36s] (i.e. [T_{89}, T_{100}]), from 100 simulations".

If a case fails with 50 simulations there is no upper confidence interval and the result should be quoted as:

"Total Evacuation Time is 89m40s (i.e. T_{48}), 99.9%CI [83m11s, N/A] (i.e. [T_{40}, N/A]), from 50 simulations"

3 COMPARITIVE DEMONSTRATION OF THE ACCURATE CI METHOD WITH THE OLD AND NEW IMO METHODOLOGIES

The main consideration in using the CICT is to ensure it is as accurate as the current IMO (2016) method in determining whether a vessel has passed or failed the assembly time criterion whilst reducing the computational burden when a design clearly passes or fails the criterion by a significant margin.

Eight different methods are compared to assess the number of true passes and true fails against the number of false passes and false fails based on two "known" distributions of simulation TETs with a "known" population 95th percentile times. The eight methods are:

1) Using 50 simulations as specified in the old IMO (2007) guidelines. (50 in Fig. 9 and Fig. 10).

2) Using 500 simulations as specified in the current IMO (2016) guidelines. (500 in Fig. 9 and Fig. 10)

3) Using the convergent method (see Eq. 5) suggested in the current IMO (2016) guidelines. (MSC-C in Fig. 9 and Fig. 10)

4) Using a simple convergence criterion (see Eq. 6) analogous to methods used in numerical solvers (Simple Conv in Fig. 9 and Fig. 10)

5) Using the CICT with a 95% confidence interval with $n_{max} = 500$. (CI(95%) in Fig. 9 and Fig. 10)

6) Using the CICT with a 99% confidence interval with $n_{max} = 500$. (CI(99%) in Fig. 9 and Fig. 10)

7) Using the CICT with a 99.9% confidence interval with $n_{max} = 500$. (CI(99.9%) in Fig. 9 and Fig. 10)

8) Using the CICT with a 95% confidence interval with $n_{max} = 1000$. (CI(95%)-1000 in Fig. 9 and Fig. 10)

3.1 Convergent Method described in IMO (2016) MSC.1/Circ 1533 (Method 3)

The convergent method (method 3) described in IMO (2016) is briefly outlined below:

Convergence is achieved when the Eq. 6 is satisfied.

$$\left| PFCT - \tau^{mean50} \right| \geq \tau^{\max 50} - \tau^{\min 50} \qquad (6)$$

Where

$$\tau^{mean50} = \sum_{j=N-49}^{N} \tau^{j} \Big/ 50 \,,$$

$$\tau^{\max 50} = MAX_{j=N-49}^{N} \left(\tau^{j} \right),$$

$$\tau^{\min 50} = MIN_{j=N-49}^{N} \left(\tau^{j} \right),$$

The simulations are incrementally performed in batches of 50. In IMO (2016) the expression is used for calculating the convergence of the travel (and response) time e.g. excluding launch and embarkation time. However, the nature of the relation will equivalently identify convergence irrespective of whether travel time or TET is used provided the PFCT is modified appropriately.

3.2 Simple Numerical Solver Convergence Method (Method 4(a) and 4(b))

A typical convergence criterion for a variable ϕ utilised by iterative numerical solvers at the i^{th} iteration is described by eq. 7.

$$\frac{\left|\phi^{i}-\phi^{i-1}\right|}{\phi^{i}} < Tolerance \qquad (7)$$

This convergence is attractive due to its simplicity and it is tempting to apply it to random sampling (e.g. outputs from stochastic evacuation models). However, there are several key differences between numerical solvers and random sampling for a sampling statistic that make this inappropriate for the latter:

- For a numerical solver, when $\phi^{i-1} = \phi^{converged}$ then generally $\phi^{i} = \phi^{converged}$ but for sampling if $\phi^{S=i-1} = \phi^{P}$ then generally $\phi^{S=i} \neq \phi^{P}$ and convergence may not be detected.
- For a numerical solver, when $\phi^{i-1} \neq \phi^{converged}$ then generally $\phi^{i} \neq \phi^{i-1}$ but for sampling when $\phi^{S=i-1} \neq \phi^{P}$ it is quite possible that $\phi^{S=i} \approx \phi^{S=i-1}$ and therefore convergence may be incorrectly and prematurely detected.
- It is unclear how the tolerance of the convergence method relates to any form of error bar or CI for ϕ. For a numerical solver the tolerance of this criteria is generally considered to be the error bar. However, this will not be the case for sampling.

If this method for convergence is applied to a sampling statistic the incorrect premature convergence issue must be addressed. Two possible approaches to overcome this problem are investigated. The first approach (method 4a), as adopted by Ronchi et al. (2014), requires that the test must be satisfied 10 consecutive times. It is noted that only this aspect of the method proposed by Ronchi et al. (2014) for convergence testing is examined here. In addition at least 50 simulations must be performed to satisfy current IMO guidelines. Thus the second approach (method 4b) tests the difference between the sample statistics when multiples of 50 simulations have been determined (see eq. 8). Thus for this approach a minimum of 100 simulations is required.

$$\frac{\left|\tau^{S=50i}-\tau^{S=50(i-1)}\right|}{\tau^{S=50i}} \leq Tolerance \text{ , where } i \text{ is the } i^{th} \text{ batch of 50 simulations} \qquad (8)$$

The tolerance level for both method 4a and method 4b has been set to the strictest possible level (i.e. zero).

3.3 Testing Distributions

The first "known" distribution of TETs (called Distribution A) was generated by performing 10,000 assembly simulations on a cruise ship geometry consisting of 13 decks split into 7 main vertical zones (MVZ) with four assembly stations and a total passenger population of 1779. This geometry forms part of a ship evacuation model validation data set that is publically available (Galea et al., 2012, 2013, 2014). The passenger characteristics and spatial distribution are randomly re-specified for every simulation according to IMO (2016) guidelines. The maritimeEXODUS (Gwynne et al, 2003) software was used to generate the simulation TETs however any suitable stochastic evacuation model could have been used. The TET of each simulation was saved to a database of results. The TETs for the sample of 10,000 repeat simulations varied from 45m 25s to 63m 20s. The sample 95th percentile time ($\tau^{S=10000}$) which is considered to be a good estimate for the population 95th percentile TET was found to be 57m 25s with 99.9% CI [57m 11s, 57m 36s]. A computer program was written to randomly sample the database to generate the "small" samples used to test the methods. A single sample was equivalent to a set of simulations needed to establish whether the PFCT was met. This database of TETs was also used to create the τ^S distributions in Fig. 1 and Fig. 6. The approach of sampling a database of results was used as it took a computational effort of 14 CPU-core-days to produce the original 10,000 simulation sample (as each simulation took approximately 2 minutes to be computed). Had unique simulations been used for testing purposes then the computational effort to generate the one million TET values used in this analysis would have been of the order of 15,000 CPU-core-years using current computational technology. The Law of Large Numbers (Rice, 1995) would also suggest that this distribution will be a close approximation of the true population distribution.

The second known distribution of TETs (called Distribution B) was a normal distribution with the τ^P set to 57m 25s using a mean value of 47m 10s and standard deviation of 6m 14s.

The PFCT for a passenger ship with more than three MVZs is 80 minutes (4800s). The eight methods were tested against this criterion. One million samples were generated for each of the methods and the number of passes and fails recorded (see Table 5). Although 'ideally' an infinite number of samples would give the 'precise' answer, one million samples were chosen as this was computationally tractable and would result in a reasonably precise estimate of the false positives/negatives. For example, if the percentage of false positives were 10% then the CI(95%) for that value would be $\pm$ 100*1.96*($\sqrt{}$(0.1*0.9/10^6) = $\pm$ 0.06%. The average number of simulations used in each samples was also recorded. It is noted that identical results were obtained using Distribution A and B.

Table 5. Comparison of testing methods for the known population of TETs with PFCT = 80 minutes.

Method	**Average sample size**	**Passes (95th Sample < PFTC)**	**Fails (95th Sample > PFTC)**
1) MSC.1/Circ 1238	50	100%	0%
2) MSC.1/Circ 1533	500	100%	0%
3) MSC-C	50	100%	0%
4a) Simple Conv	56	100%	0%
4b) Simple Conv	261	100%	0%
5) CI (95%)	100	100%	0%
6) CI (99%)	150	100%	0%
7) CI (99.9%)	150	100%	0%
8) CI (95%)-1000	100	100%	0%

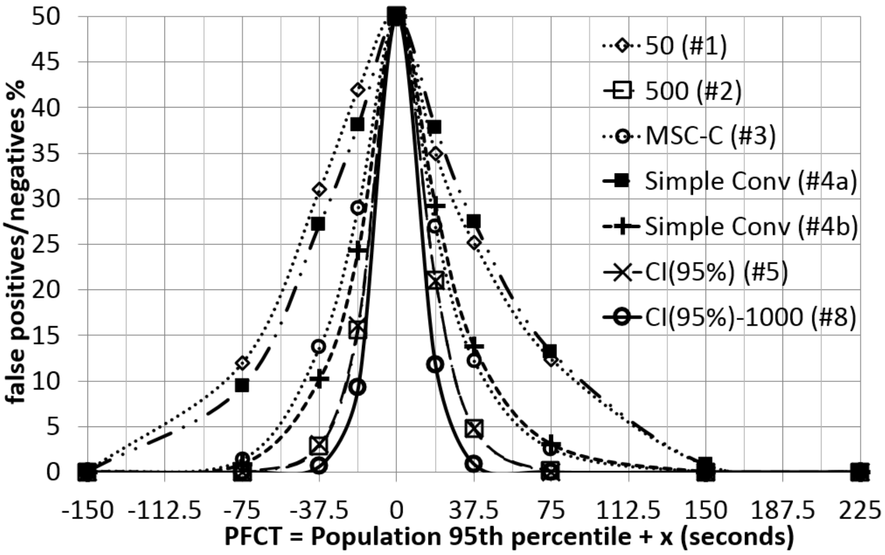
The design easily passes the 80 minute criterion and required the minimum sample size for the CICTs and method 3. This was expected as the "known" τ^P was 57m 25s (3445s) with 99.9% CI [57m 11s, 57m 36s] is significantly less than the 80 minute criterion. Referring to Table 4 the $R_{CI(95\%)}$ is 253.4s for a 100 simulation sample and the |PFCT - τ^P| is 1355s. The ratio of (PFCT - τ^P) to $R_{CI(95\%)}$ is 5.3, well beyond the maximum quoted ratio, 3/2, in Table 3 and indicates the pass would be easily determined by a sample of 100 simulations. Method 4 is independent of the PFCT so requires more simulations than the other methods in this instance and therefore uses far more simulations than is necessary to accurately determine whether the design has passed.

In order to fully assess the various methods τ^P should be close to the PFCT. From Table 4 it is known the average CI (95%) range for a 500 simulation sample is 88.2s. The methodologies were again tested this time assuming PFCT times that were 150s, 75s, 37.5s and 18.75s less than the known τ^P and PFCT times that were 225s, 150s, 75s, 37.5s and 18.75s greater than the known τ^P. One million samples were generated for each methodology for each PFCT time. Ideally it would be expected that all the samples with PFCTs less than τ^P would fail and all the samples with PFCT above τ^P would pass. In general there will be an increasing number of false positives and false negatives as the PFCT gets closer to τ^P. This testing was performed using Distribution A and Distribution B.

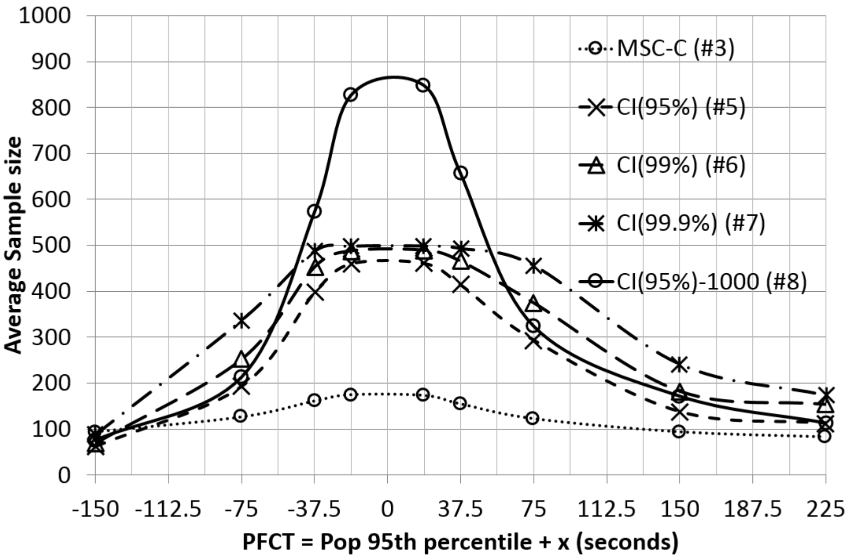
3.4 Testing using the modified PFCTs and Distribution A

The percentage of false positives/negatives for each PFCT for the eight methods applied to TET Distribution A is depicted in Fig. 9. The percentage of correctly determined passes/fails for each PFCT is (100 – percentage of false negatives/positives).

It can be seen from Fig. 9(a) that the 50 simulation requirement specified in IMO (2007) Guidelines is more susceptible to false positives or false negatives than the other methods. Furthermore, increasing the minimum number of simulations from 50 to 500, as specified in the current IMO (2016) Guidelines clearly reduces the number of false positives. At PFCT = τ^P – 37.5s the specifications in the old IMO (2007) Guidelines result in 31±0.09% false positives compared to just 3±0.03% for the 500 simulation limit specified in the current IMO (2016) Guidelines. The CI(95%) (#5) method (n_{max} = 500) is essentially as accurate as the full 500 simulations over the range of PFCTs as the plotted data points very closely match the full 500 (#2) simulation samples. The CI(99%) (#6) and CI(99.9%) (#7) are even closer to the full 500 simulation results and as they lie between the CI(95%) (#5) results and the full 500 (#2) simulation samples are not illustrated in Fig 9(a) as they would be indistinguishable from those plots. The number of false positives or false negatives caused by the CICT methods is very small and equivalent to the number of false positives or false negatives caused by limiting the sample size to a maximum of 500 simulations. The number of false positives/negatives produced by the CICT is further reduced for the CI(95%) (N_{MAX} = 1000) method. Indeed, this method has a superior pass/fail accuracy compared to all the other methods illustrated. The MSC-C method and the simple convergence method perform significantly better than using 50 simulations but do not perform as well as 500 simulations or the CICTs at reducing the number of false positives/negatives. It should be noted that the simple convergence methods even with the tolerance set to zero perform poorly compared the CICTs with method 4a being particularly poor.



(a)



(b)

Fig. 9. Plot of false positives/negatives for a range of PFCTs using 7 different assessment methods (a) and the average sample size for 5 assessment methods (b) (Distribution A).

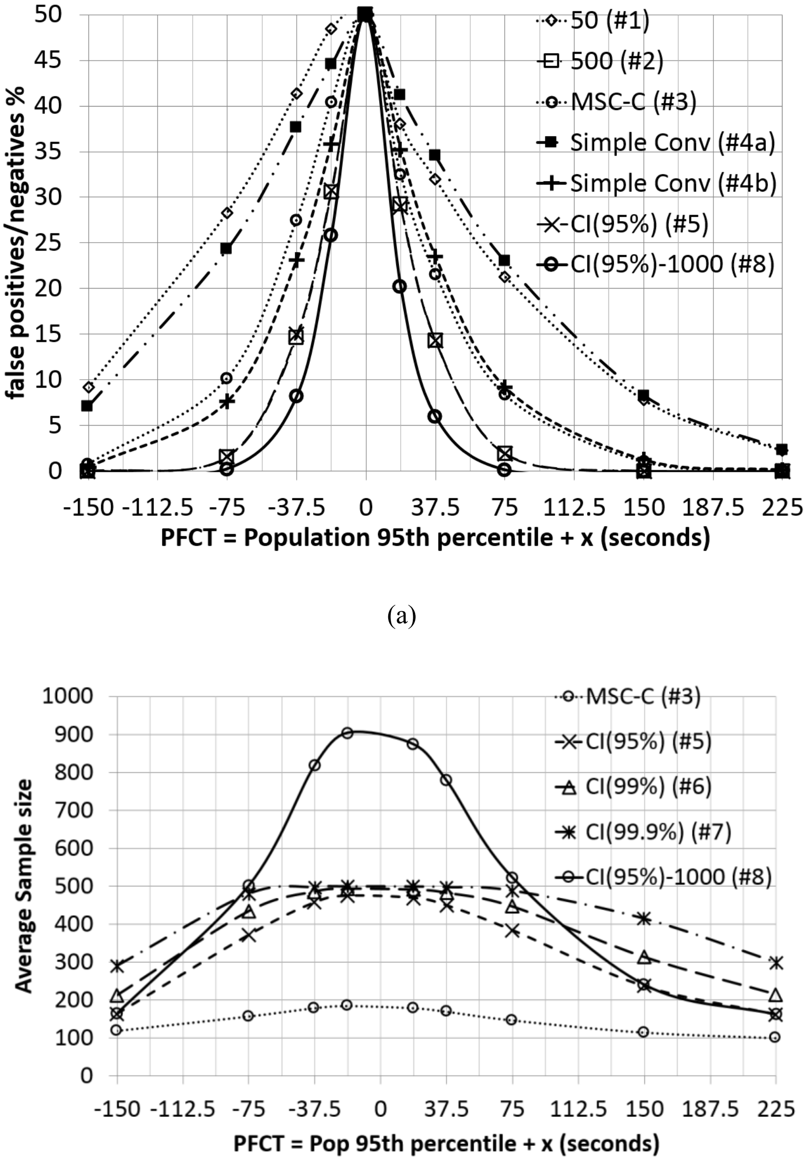
When the PFCT is equal to τ^P it is expected that 50% of the samples would pass the criterion and 50% would fail the criterion but it cannot strictly be identified as a false positive or a false negative, the point is identified to aid curve plotting.

The average sample size (number of simulations) required by each PFCT for each method for Distribution A is depicted in Fig. 9(b). As the simple convergence methods (4a and 4b) are independent of proximity of PFCT to τ^P the average number of cases were found to be 56.6 and 261 respectively. When CICTs are used, significantly fewer simulations than 500, on average, are required to establish whether the design has passed or failed when the PFCT is “far” from the τ^P. The CI (95%) method requires far fewer simulations than the other CICTs but with no significant effect on the overall accuracy (see Fig. 9). For example the average number of simulations required for a PFCT = (τ^P - 75s) is 193 for CI(95%), 252 for CI(99%), and 337 for CI(99.9%). The proportion of false negatives is only 0.01±0.002% for CI(95%), 0.003±0.001% for CI(99%), 0.003±0.001% for CI(99.9%) and 0.002±0.0009% for 500 simulations. The CI(99.9%) requires 75% more simulations than CI(95%) and 500 simulations requires 159% more simulations than CI(95%) with a negligible increase in pass/fail accuracy (see Fig. 9). There is no significant difference in pass/fail accuracy between the full 500 simulations and the CI(99%) and CI(99.9%) methods. Increasing n_{max} to 1000 clearly decreases the number of false positives/negatives near the τ^P but significantly more simulations are required to improve this accuracy. However, for a PFCT of (τ^P - 75s) it can be seen that an average of 573 simulations are required for CI(95%)-Nmax=1000 but the number of false positives is only 0.63±0.02% compared to 2.83±0.03% obtained with 500 simulations and 3.0±0.03% obtained with CI(95%)-Nmax=500.

It can be seen that the MSC-C method (Fig. 9(b)) also tends to increase the sample size required as PFCT tends to τ^P, although not to the same extent as the CICTs, but the CICTs have superior pass/fail accuracy.

3.5 Testing using the modified PFCTs and Distribution B

As was observed with Distribution A (see Fig. 9(a)) the 50 simulation specification required by the old IMO (2007) Guidelines is more susceptible to false positives or false negatives than the other methods when applied to Distribution B (see Fig. 10(a)). At PFCT = (τ^P – 75) it can be seen that the 50 simulation specification of the old IMO Guidelines leads to 28±0.09% false positives compared to 1.5±0.02% with 500 simulations (as specified in the current IMO (2016) Guidelines) and 0.3±0.01% with the CI(95%)-1000. The CI(95%) (#5) method (n_{max} = 500) is essentially as accurate as the full 500 simulations over the range of PFCTs as the plotted data points very closely match the full 500 (#2) simulation samples. The CI(99%) (#6) and CI(99.9%) (#7) are even closer to the full 500 simulation results and as they lie between the CI(95%) (#5) results and the full 500 (#2) simulation samples are not illustrated in Fig 10(a) as they would be indistinguishable from those plots. The number of false positives or false negatives caused by the CICT methods is very small and equivalent to the number of false positives or false negatives caused by limiting the sample size to a maximum of 500 simulations. The number of false positives/negatives produced by the CICT is further reduced for the CI(95%) (N_{MAX} = 1000) method. Indeed, this method has a superior pass/fail accuracy compared to all the other methods illustrated. The simple convergence methods again perform poorly compared to the CICTs with method 4a being particularly poor.



(a)

(b)

Fig. 10. Plot of false positives/negatives for a range of PFCTs using 7 different assessment methods (a) and the average sample size for 5 assessment methods (b) (Distribution B).

The average sample size (number of simulations) required for each PFCT for each method for Distribution B is depicted in Fig. 10(b). As the simple convergence methods (4a and 4b) are independent of proximity of PFCT to τ^P the average number of cases was found to be 56.7 and 265 respectively. As was noted for Distribution A significantly fewer simulations than 500, on average, are required to establish whether design has passed or failed when the PFCT is “far” from τ^P. The CI (95%) method requires far fewer simulations than the other

CICTs but with no significant effect on the overall accuracy (see Fig. 10a). For example the average number of simulations required for a PFCT = τ^P – 37.5s is 295 for CI(95%), 369 for CI(99%), and 443 for CI(99.9%). The proportion of false negatives is 0.47±0.01% for CI(95%), 0.22±0.009% for CI(99%), 0.2±0.009% for CI(99.9%) and 0.2±0.009% for 500 simulations. The CI(99.9%) method requires 50% more simulations than CI(95%) and 500 simulations requires 69% more simulations than CI(95%) with a negligible increase in pass/fail accuracy (see Fig. 10). In addition the minimum number of simulations for the CI (95%) method is 100 simulations compared to 150 simulations for the CI(99%) and CI(99.9%) method. A "large" number of simulations are only required when the PFCT is "close" to τ^P which is identified by the CICT methodology.

4 DISCUSSION OF CONVERGENCE MEASURES

From the above analysis it is clear that CICT method has a superior accuracy to the other tested measures of convergence for τ. Here we explore why the CICT approach is superior to the other tested methods using one set of 500 simulations of the previous cruise ship example. As this analysis relates to just one possible sample set, the issues identified may not relate to all potential candidate sample sets however, the identified issues are expected to frequently arise and hence impact the performance of the various approaches.

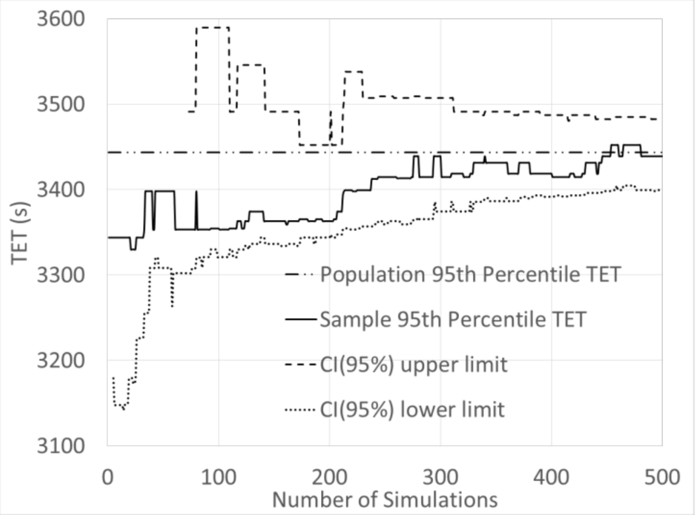


Fig. 11. Plot of variation of τ^S and the CI(95%) with number of simulations in a single sample.

In this case, the "population" distribution consists of 10,000 simulations and $\tau^P = \tau^{S=10000}$ = 3444s. Furthermore, we will consider for demonstration purposes the PFCT is set to 3400s, so the design should fail as τ^P = 3444s. Presented in Fig. 11 is τ^P (dash-dot-dot line), τ^S (solid line) and the limits of the CI(95%) (the upper limit is represented by the dashed line and the lower limit is represented by the dotted line) as a function of increasing sample size.

Basing a decision on only the first 50 simulations (method 1) would result in $\tau^{s=50}$ = 3366s and would lead to a false positive with the difference between τ^P and $\tau^{s=50}$ is 45s (3443s-3398s). Basing the decision on the full 500 simulations (method 2) would result in $\tau^{S=500}$ = 3439s and would lead to a fail. In this example, it is fortuitous that τ^S is very close to τ^P and as a result the correct conclusion is drawn i.e. that the vessel fails. However, in reality the difference between τ^S and τ^P could be quite large increasing the chances of an incorrect decision being made. Using the current IMO (2016) Guidelines, there is no way of knowing how reliable the decision based on 500 simulations may be. This can easily be rectified by specifying the CI associated with the estimation of τ^S (using the information in Table 1). The 95% CI for the sample of 500 simulations is [3399s, 3482s] which indicates there is some uncertainty that τ^P is greater than the PFCT, as the PFCT lies within the CI, at a 95% confidence level. This suggests that another sample of 500 simulations may have suggested that the design could pass. This indicates that the analysis based on 500 simulations, while in this case indicating the correct result (i.e. a fail) is unreliable and a greater sample size should be used to resolve the situation.

From Fig. 11 it is noted that τ^S can be relatively stable over a significant number of simulations even though it is not very close to τ^P. When applying the simple convergence method 4b, $\tau^{s=150} = \tau^{s=200}$ = 3363s and so using this method the population 95th percentile time is predicted to be 3363s. Thus the converged sample 95th percentile time leads to a false positive result. Method 4a predicts convergence after 58 simulations have been performed with $\tau^{S=58}$ = 3398s and again results in a false positive. It is clear that the simple convergence methods, even if adapted to address the statistical nature of the predicted parameter, are not reliable estimators of convergence when applied to τ. While the failure of the simple convergence methods are demonstrated here for τ it is argued that the technique is similarly flawed for other parameters as the observations in section 3.2 would apply to all sampling statistics. Simple convergence methods and the variants described in this paper are therefore considered to be unreliable for demonstrating convergence for stochastic evacuation models.

Using CICT (method 5) the design is not considered to be a clear pass or fail because the PFCT lies within the CI (i.e. 95% CI[3399s, 3482s]) and therefore τ^P could be higher or lower than the PFCT. In this case as 500 simulations have been performed we would compare the predicted sample 95th percentile time τ^S, with the PFCT. As found in method 2, $\tau^{S=500}$ = 3439s and the design is deemed to fail which is consistent with τ^P. However, as the PFCT lies within the CI there is possibility that the analysis could have resulted in a pass. Using CICT with a larger maximum number of simulations within the sample (method 8) resolves this situation by reducing the size of the CI resulting in a fail determination in 550 simulations with TET of 3452.3s with 95%CI [3409s, 3482s]. When method 3 (see eq. 6) is used, a pass is determined after 150 simulations as |3400s

– 3362.6s| > (3374.4s – 3353.3s), leading to a false positive with the difference between τ^P and τ^{mean} being 81s (3443s-3362s).

5 SUGGESTED USE OF THE CICT AND ITS VARIANTS

The CICT has been designed for easy manual application by running the simulations in batches of 50. However, if the method is automated within an egress model it then becomes practical, although not necessary, to check the convergence after every single simulation to further reduce the number of simulations that are required. The minimum number of simulations required to potentially determine a pass is then 72 for CI(95%), 104 for CI(99%), and 149 for CI(99.9%). The minimum number of simulations to potentially determine a fail is 3 for CI(95%), 4 for CI(99%), and 4 for CI(99.9%). However, to be compliant with the current IMO (2016) Guidelines, when using a convergence method a minimum of 50 simulations is required.

The CICT method will ideally result in identifying a clear pass or fail prior to the maximum number of simulations being necessary. However, in some situations, the CICT method will not return a clear pass or fail before the maximum number of simulations is reached because the PFCT lies within the CI. In such circumstances there are several possible ways to proceed.

1. To maintain consistency with the current IMO (2016) Guidelines, the pass/fail is determined by directly comparing $\tau^{S=500}$ to the PFCT.
2. Define a new upper limit to the number of simulations (e.g. 1000) and continue with the testing process until a pass/fail result is obtained. Should the new maximum number of simulations be reached and the result is still uncertain there are three options:
    a. define a new upper limit and continue testing however, as the convergence is slow ($1/\sqrt{n}$) convergence will be slow beyond a few thousand simulations,
    b. determine a pass/fail based on directly comparing τ^S to the PFCT (as in option 1) or
    c. declare that the design has failed as the τ^P is too close to the PFCT.
3. Declare that the design has failed as the τ^P is too close to the PFCT.

The CICT has been applied to determining whether τ^P is greater than or less than a PFCT. However, the CICT can be applied to any simulation output parameter, not just the TET, for example the predicted number of fatalities or a measure of congestion. Furthermore, the CICT could be easily adapted for any predicted *population parameter* (here referring to statistical measures), such as the *population* mean or standard deviation, provided that a CI can be defined, to determine whether it meets a pass/fail criterion. However, it is noted that while it is possible to analytically define a CI for the population mean and 95th percentile parameters, it may not be possible for some parameters. Nevertheless, where no analytical CI exists for a parameter it may be possible to bootstrap (Efron and Tibishirani, 1993) a computational CI.

While the CICT method has been developed to address the requirements of the IMO Guidelines, it can be applied to any situation in which a simulated parameter is being compared against a pass/fill criterion. In the building industry the pass/fail criterion could be the Available Safe Egress Time (ASET) determined by fire modelling calculations as part of a traditional Required Safe Egress Time (RSET) < ASET calculation. In the aviation industry, the pass/fail criterion could be the regulatory specified 90s for passengers to evacuate the aircraft (FAR, 1999).

Although not demonstrated in this paper it could be argued that the so-called "5σ" (99.99997%) confidence level should be used rather than the 95% confidence level. The CICT can be easily modified to address this requirement and would require a minimum of 307 simulations to potentially define a pass and a minimum of 7 simulations to potentially define a fail. However, using a "5σ" confidence level would not significantly improve the pass/fail determination accuracy but would lead to a very large increase in the number of simulations that would be needed to determine whether a design/scenario had passed or failed.

Finally, while the CICT has been specifically designed to determine whether τ^P is greater than or less than a particular critical value, the methodology can be adapted to determine the value of a parameter to a given level of precision at a specified confidence level. Rather than checking the upper and lower CI bounds against a PFCT, the R_{CI} can be compared to an absolute or relative extent and will be "converged" when R_{CI} is less than the stipulated value (see eq. 9). However, as the convergence is slow (($R_{CI} \propto 1/\sqrt{n}$) → 0 as $n \to \infty$) it may not always be possible to achieve the desired tolerance and so in such circumstances it is necessary to report the absolute or relative tolerance achieved at some maximum number of simulations.

$$\frac{R_{CI}^{S=i}}{\phi^{S=i}} < Tolerance \tag{9}$$

7 CONCLUSIONS

Current international guidelines, specified by the International Maritime Organization (IMO), for determining representative assembly times for large passenger ships using evacuation simulation models utilise a brute force method requiring the generation of 500 evacuation simulations. The representative assembly time for the vessel design for a particular evacuation scenario is considered to be the 95th percentile assembly time generated from the series of 500 simulations. This time is intended to be a good approximation to the actual 95th percentile time that would be generated from the population of all possible permitted random permutations of the simulation parameters and random behavioural permutations of the simulated agents. This approach has two significant deficiencies. The first is that an excessive computational effort may be required to determine the representative time as in some cases it may be possible that the sequence of predicted representative assembly times is sufficiently converged to determine whether the design has passed or failed in significantly fewer than 500 simulations.

To address the first issue the IMO guidelines also specify an alternative approach, the so-called convergent method, to determine if the predicted 95th percentile assembly time has converged to the required time in fewer than 500 simulations. However, as demonstrated in this paper the convergent method presented in the IMO guidelines has a significant propensity to produce false positive and false negative determinations. This is because the method settles on a 'converged' representative time which is a poor approximation of the actual representative time. Indeed, it was demonstrated that the brute force method, while requiring significantly more simulations than the IMO convergent methodology, produced significantly fewer false positive/negative results.

The second more significant issue with the IMO Guidelines is that the sequence of predicted representative assembly times may require more than 500 simulations to accurately produce a good representation of the actual 95th percentile assembly time for the entire population of cases. Thus it is possible for the brute force method to produce false positive or false negative results. This is because the brute force method settles on a 'converged' representative time which is a poor approximation of the actual representative time. This is considered a very serious limitation of the current approach as it can lead to a predicted representative time which is either larger than the pass/fail criteria time (PFCT) producing a false negative result i.e. predicting a failure when in fact the actual time is less than the PFCT or of greater significance, producing an approximation which is smaller than the PFCT generating a false positive result i.e. predicting a pass when in fact the actual time is greater than the PFCT.

To address the failings of both methods proposed in the IMO Guidelines the Confidence Interval Convergence Test (CICT) method was developed. The CICT was demonstrated to be as accurate as the brute force method (producing an equivalent number of false positives/negatives) but requiring far fewer simulations, saving computation time and costs, thereby addressing the first issue. Furthermore, the CICT was significantly more accurate than the IMO convergent method albeit requiring more simulations. Furthermore, for the CICT method, when the scenario is a clear fail it only requires 50 simulations to demonstrate and if the case is a clear pass it only requires 100 simulations. Large numbers of simulations are generally only required if the PFCT is "close" to the 95th percentile of the actual population of cases.

The CICT methodology is straightforward to implement and reverts to the brute force methodology as specified in the current IMO Guidelines when 500 simulations are required. Used in this way the CICT method has comparable accuracy to the brute force method presented in the IMO Guidelines. To address the more significant second failing of the brute force method, the CICT method can achieve superior overall accuracy (i.e. predicts fewer false positives/negatives) to the brute force method if n_{max} is set to 1000 and will generally use less than 500 simulations in practice. Furthermore, as the CICT method produces a CI together with the sample 95th percentile time, there is always an unambiguous indication when the predicted result may be in doubt, thereby highlighting the need to consider additional simulations before a final pass/fail decision can be made. As the methods identified within the current IMO Guidelines do not require the determination of a CI, using these methods it is not possible to identify when an uncertain determination is made.

It has also been demonstrated that simple convergence methods, as used for iterative numerical solvers, should not be used for the determination of 95th percentile values as they are subject to producing a significant number of false positive/negative results even when corrective measures are taken to limit premature convergence. Furthermore, it is suggested that their use for other sampling statistics is also questionable and should not be used to demonstrate convergence of any stochastic evacuation model – including those used in the building industry.

Finally, it is also recommended that all TETs are quoted to include the highest possible confidence interval with the sample size used and that population parameters (walking speeds, response times, etc) and initial locations are randomised within the limits prescribed for every simulation within a sample to avoid potential biasing. While it is possible to use the CICT method manually it is envisaged that the methodology will be automated by embedding it within the evacuation modelling tool.

Acknowledgements

This project received funding from the University of Greenwich REF Development Grant.

References

Agresti, A., Coull, B. A., 1998, Approximate is better than 'exact' for interval estimation of binomial proportions, The American Statistician. 52: 119–126. doi:10.2307/2685469

Averill, J. D., 2011, Five grand challenges in pedestrian and evacuation dynamics, Proceedings of the 5th international pedestrian and evacuation dynamics conference, pp1-11, 2011, http://dx.doi.org/10.1007/978-1-4419-9725-8_1

Brown, L. D., Cai, T. T., DasGupta, A., 2001, Interval Estimation for a Binomial Proportion. Statist. Sci. 16 (2001), no. 2, 101--133. doi:10.1214/ss/1009213286.

Clopper, C. J., Pearson, E. S., 1934, The Use Of Confidence Or Fiducial Limits Illustrated In The Case Of The Binomial, Biometrika (1934) 26 (4): 404-413 http://dx.doi.org/10.1093/biomet/26.4.404

Efron, B., Tibshirani, R., 1993, An Introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall/CRC., 1993, ISBN 0-412-04231-2

FAR, 1999, Part 25.Appendix J Airworthiness Standards: Transport Category Airplanes. Including amendment 25-98 as published in the Federal Register on February 8th, 1999, Washington DC, USA.

Galea, E.R., 2006, Proposed Methodology for the Use of Computer Simulation to Enhance Aircraft Evacuation Certification, AIAA Journal of Aircraft, Vol 43, Number 5, pp 1405-1413, 2006. DOI: 10.2514/1.20937

Galea, E. R., Togher, M., and Lawrence, P., 2010, Investigating the impact of Aircraft Exit Availability on Egress Time Using Computer Simulation, Proceedings Pedestrian and Evacuation Dynamics 2008, Ed: Kligsch, W. W. F., et al, Springer-Verlag Berlin Heidelberg, ISBN 978-3-642-04503-5, DOI:10.1007/978-3-642-04504-2_35, pp411-424, 2010

Galea, E. R., Filippidis, L., Deere, S., Brown, R., Nicholls, I., Hifi, Y., Besnard, N., 2012, IMO Information paper - The SAFEGUARD validation data-set and recommendations to IMO to update MSC/Circ. 1238, SAFEGUARD Passenger Evacuation Seminar 30 November 2012, London, UK, pp. 98-103, ISBN 978-1-909024-08-3, 2012

Galea E.R., Deere, S., Brown, R., Filippidis, L., 2013, An Experimental Validation of an Evacuation Model using Data-Sets generated from Two large Passenger Ships, The Society of Naval Architects and Marine Engineers (SNAME) Journal of Ship Research, Vol 57, number 3, pp155-170, Sept 2013. http://dx.doi.org/10.5957/JOSR.57.3.120037

Galea, E.R., Deere S., Brown R. and Filippidis, L., 2014, A Validation Data-Set and Suggested Validation Protocol for Ship Evacuation Models. Fire Safety Science, Proceedings of the 11th International Symposium, IAFSS, 2014, pp. 1115-1128, http://dx.doi.org/10.3801/IAFSS.FSS.11-1115

Gwynne, S, Galea, E.R., Owen, M., Lawrence P.J. and Filippidis L., 1999, A Review of the Methodologies Used in Evacuation Modeling, Fire and Materials, v23, 6, pp383-389, Nov-Dec 1999. http://dx.doi.org/10.1016/s0360-1323(98)00057-2

Gwynne, S., Galea, E. R., Lawrence, P. J. and Filippidis, L., 2001, Modelling Occupant interaction with fire conditions using the buildingEXODUS evacuation model, Fire Safety Journal, 36, pp 327-357, 2001. http://dx.doi.org/10.1016/S0379-7112(00)00060-6

Gwynne, S., Galea, E.R., Lyster, C., Glen, I., 2003, Analysing the Evacuation Procedures Employed on a Thames Passenger Boat using the maritimeEXODUS Evacuation Model, Fire Technology, Vol 39, No. 3, pp. 225-246, 2003, http://dx.doi.org/10.1023/A:1024189414319

Ha S, Ku NK, Roh MI, Lee KY., 2012, Cell-based evacuation simulation considering human behavior in a passenger ship. Ocean Eng 2012; 53: 138-152. http://doi.org/10.1016/j.oceaneng.2012.05.019

IMO, 2007, Guidelines for Evacuation Analysis for New and Existing Passenger Ships, IMO MSC.1/Circ 1238, 30 Oct 2007.

IMO, 2016, Revised Guidelines for Evacuation Analysis for New and Existing Passenger Ships, IMO MSC.1/Circ 1533, 6 June 2016.

Kirkpatrick, S., and E. Stoll, 1981; “A Very Fast Shift-Register Sequence Random Number Generator”, Journal of Computational Physics, V. 40. pp. 517-526 DOI: 10.1016/0021-9991(81)90227-8

Korhonen, T., Hostikka, S., Heliovaara, S., Ethamo, H., 2008, FDS+Evac: an agent based fire evacuation model, Pedestrian and Evacuation Dynamics, Springer, Wuppertal, Germany, pp. 109–120, 2008. http://dx.doi.org/10.1007/978-3-642-04504-2_8

Kuligowski, E.D., Peacock, R.D., and Hoskins, B. L., 2010, A Review of Building Evacuation Models, 2nd Edition, National Institute of Standards and Technology (NIST) Technical Note 1680 November 2010.

Matsumoto, M., Nishimura, T., 1998, Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator, ACM Transactions on Modeling and Computer Simulation. 8 (1): 3–30. doi:10.1145/272991.272995

Meacham, B.,Lord, J., Moore, A., Fahy, R., Proulx, G., Notarianni, K., 2004, Investigation of Uncertainty in Egress Models and Data, Proceedings of the 3rd International Symposium on Human Behaviour in Fire, Belfast, UK, September 01-03, 2004, pp. 419-428, (NRCC-47308)

Meyer-König, T., Valanto, P., Povel, D., 2005, Implementing Ship Motion in AENEAS — Model Development and First Results, Pedestrian and Evacuation Dynamics 2005, pp429-441, http://dx.doi.org/10.1007/978-3-540-47064-9_41

Neyman, J., 1937, Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability, Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences 236.767 (1937): 333–380. http://dx.doi.org/10.1098/rsta.1937.0005

Papoulis, A., 1984, Bernoulli Trials, Probability, Random Variables, and Stochastic Processes (2nd ed.). New York: McGraw-Hill. (1984) pp. 57–63

Park, JH., Lee, DK., Kim, HT., Yang, YS., 2004, Development of evacuation model for human safety in maritime casualty, Ocean engineering. 31(11), pp.1537-1547, 2004, http://dx.doi.org/10.1016/j.oceaneng.2003.12.011

Peacock, R. D, Reneke, P. A, Davis, W. D, Jones, W. W., 1999, Quantifying fire model evaluation using functional analysis, Fire Safety Journal, Volume 33, Issue 3, October 1999, pp167-184, http://dx.doi.org/10.1016/S0379-7112(99)00029-6

Pradillon, J.Y., 2003, ODIGO: A Crowd Movement Simulation Tool for Passenger Vessels, 2nd International Conference on Computer and IT Applications in the Maritime Industries, COMPIT, Hamburg, Germany, pp 108-117, 2003.

Rice, J., 1995, Mathematical Statistics and Data Analysis, 2nd ed, 1995, Duxbury Press. ISBN 0-534-20934-3

Ronchi, E., Reneke, P. A., Peacock, R. D., 2014, A Method for the Analysis of Behavioural Uncertainty in Evacuation Modelling, Fire Technology, Volume 50, pp 1545-1571, 2014, http://dx.doi.org/10.1007/s10694-013-0352-7

Thompson, P. A., Marchant, E. W, 1995, A computer model for the evacuation of large building populations, Fire Safety Journal, Volume 24, Issue 2, 1995, Pages 131-148, ISSN 0379-7112, http://dx.doi.org/10.1016/0379-7112(95)00019-P

Vassalos, D., Kim, H., Christiansen, G., Majumder, J., 2002, 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

Wilson, E. B., 1927, Probable inference, the law of succession, and statistical inference, Journal of the American Statistical Association. 22: 209–212. doi:10.1080/01621459.1927.10502953