24th International Congress on Modelling and Simulation, Sydney, NSW, Australia, 5 to 10 December 2021
mssanz.org.au/modsim2021

# Factor screening for binary responses from combat simulations

**H. Demirhan** [a], **S. Georgiou** [a], **A. Gill** [b] **and S. Stylianou** [a]

[a] *School of Science, Royal Melbourne Institute of Technology, Melbourne, VIC 3001, Australia*
[b] *Defence Science and Technology Group, PO Box 1500, Third Avenue, Edinburgh, SA 5111, Australia*
*Email: andrew.gill@dst.defence.gov.au*

**Abstract:** Combat simulations are an increasingly important tool for analysts in developing advice to Defence decision-makers around military capability, by providing an alternative to physical trials and military exercises where it may be unsafe or prohibitively expensive to operate. Anecdotally, high-fidelity simulations often follow the Pareto principle whereby a large proportion of the output is governed by a small proportion of the inputs. Thus, screening for the 'critical few from the trivially many' prior to meta-model construction is a prudent first step. Somewhat surprisingly, our literature search for binary response simulation factor screening only produced the thesis by Li (2011).

In this paper we analytically critique the hybrid screening method of Li (2011) before arguing that the original approach of Shen et al. (2010) for continuous responses can still be utilised for binary responses. A comparison of the numerical effectiveness of an apparently D-optimal design proposed by Yang et al. (2011) with a traditional factorial is performed, before exploring the sensitivity of the original Shen et al. (2010) approach to its user-specified parameters.

We show that the requirement of the modified approach of Li (2011) in using the sample covariance matrix of the regression estimates makes it susceptible to singularity issues due to rank deficiency. In contrast, we note that the critical point in the applicability of the original Shen et al. (2010) technique is the normality of the model parameters, which can follow from either the normality of the data or, importantly, any other asymptotic property of the estimation method used for model fitting. As the maximum likelihood estimates of parameters of the logistic regression model follow an asymptotically normal distribution, the Shen and Wan (2009) approach remains appropriate. We note that there is no need to assume normality of log-odds ratios such as made by Li (2011).

Numerical experiments found that the original Shen et al. (2010) method can work well when the required parameters are specified correctly. It worked efficiently with both the typical fractional factorial design and the alternative D-optimal design specifically for binary responses. As expected the D-optimal design outperforms the fractional factorial design, but the differences were relatively minor and the method seemed to behave well under both designs.

The correct specification of the user parameters is related to the problem being studied, but is influenced by parameters not under the researcher's control, for example the correct assumption of the fitted model and the identification of the variability of the true active/inactive coefficients. The method was not found to be robust to misspecification of these user parameters and the results in such cases might be misleading at best. This disadvantage makes its use highly problematic in practical situations where prior knowledge of the true underling model and range of significant/insignificant coefficients might be limited.

Further research is thus required in identifying a robust factor screening method for binary responses from stochastic simulations. Similarly, for response variables that are discrete but bounded, such as the number of combat assets lost or remaining, an effective screening approach remains elusive. Such generalisations are not straightforward and more research is needed in these directions.

***Keywords:*** *Logistic regression, Fractional factorial design, D-optimal design, simulations, screening procedures*

## 1 INTRODUCTION

Combat simulations are an increasingly important tool for analysts in developing advice to Defence decision-makers around military capability, by providing an alternative to physical trials and military exercises where it may be unsafe or prohibitively expensive to operate. However, the sophistication and fidelity of most modern combat models has the unfortunate consequence of rendering them more akin to a 'black-box' and thus of little analytical benefit by themselves. Simulation analytics is the process of fitting and exploiting an analytic meta-model of the input-output relationship between data of a simulation generated through a designed experiment (Gill et al., 2018). This 'white-box' approximation can then be used to characterise the marginal contribution to simulation outcomes from the various input factors, thus enabling either trade-space or optimisation analyses to be performed. Anecdotally, high-fidelity simulations often follow the Pareto principle whereby a large proportion of the output is governed by a small proportion of the inputs. Thus, screening for the 'critical few from the trivially many' prior to meta-model construction is a prudent first step.

Controlling the Type - I (false positive) and Type - II (false negative) errors and violation of the assumption of homogeneity of variances are important considerations for factor screening of stochastic simulation experiments (Shen and Wan, 2009). To address these issues Wan et al. (2003) and Wan and Ankenman (2006) proposed Controlled Sequential Bifurcation (CSB) where groups of factors are tested, and Shen and Wan (2009) proposed Controlled Sequential Factorial Design (CSFD) that combines sequential hypothesis testing with a factorial design. According to Shen et al. (2010), power control of CSFD is stronger than that of CSB, so they combined CSB and CSFD into a more efficient, two-staged hybrid method that exploited the complementary nature of both. Flowcharts for the implementation of CSFD, CSB and the hybrid method are provided in panels (a), (b) and (c) of Figure 1, respectively. Specific mathematical details are given in the listed references.

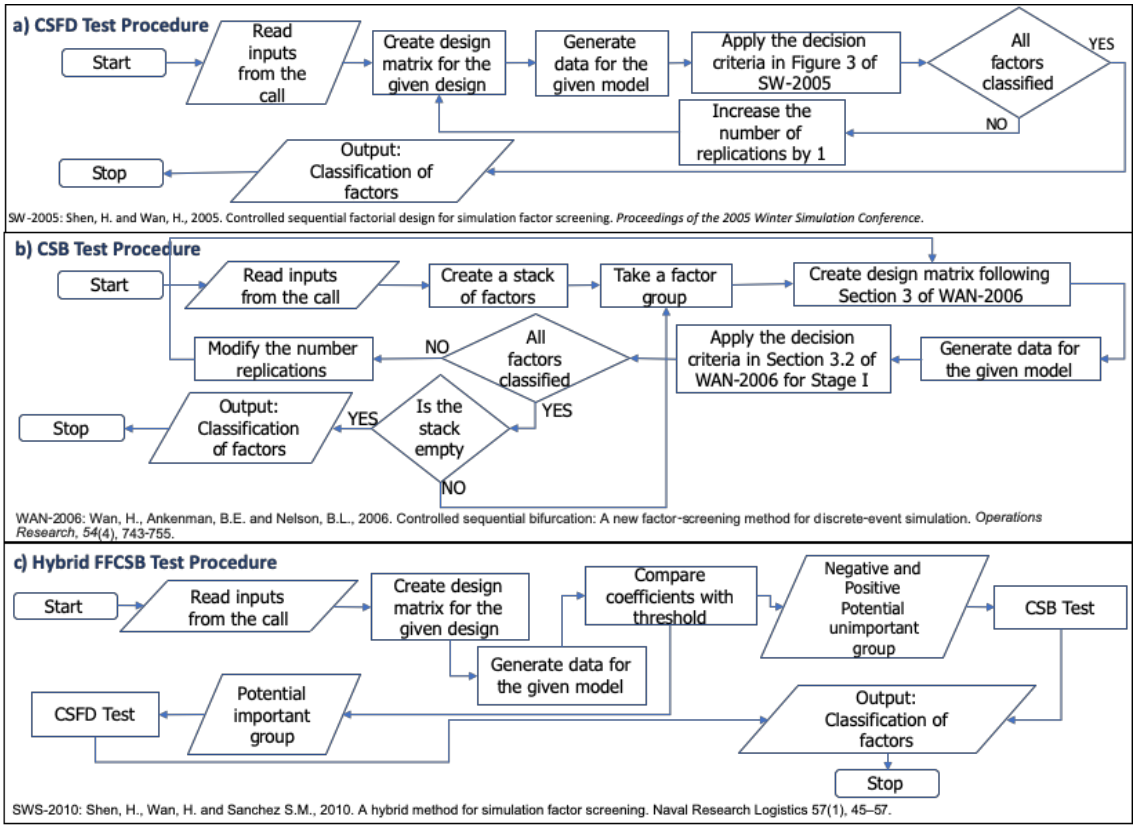


**Figure 1**. Flow charts for the implementation of (a) CSFD, (b) CSB, and (c) hybrid methods.

In all of these methods, the dependent variable $Y$ is assumed to be measured on the continuous scale. However, in practice, having binary responses in a stochastic simulation is also common (especially for combat

simulations where mission success is of primary importance). Somewhat surprisingly, our literature search for binary response simulation factor screening only produced the thesis by Li (2011), who replaces CSFD in the above hybrid method with an alternative specifically developed around binary responses. Traditionally, 2-level factorial designs are used for screening, as they are economical to construct and optimal for linear models where the focus is on main effects (Fedorov, 1972) and are used in Li (2011). However, for simulations where the response is not continuous, these designs are no longer optimal. Of particular interest then is Yang et al. (2011), who provided explicit formulae for D-optimal designs for logistic regression. In this paper we analytically critique the hybrid screening method of Li (2011) before arguing that the original CSFD can actually still be utilised for binary responses. A comparison of the numerical effectiveness of the Yang et al. (2011) design with the traditional factorial is performed, before exploring the sensitivity of the original Shen et al. (2010) approach to its user-specified parameters.

## 2 SIMULATION FACTOR SCREENING FOR BINARY RESPONSES

Defining the random variable $Y_i \in \{0, 1\}$ to represent the binary response from the simulation execution at the $i-$th design point, defined by $K$ input factors, then $Y_i \sim$ Bernoulli$(p_i)$ and the success probabilities $p_i$ are modelled as:

$$\log\left(\frac{\mathbf{p}}{1-\mathbf{p}}\right) = \boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta} \tag{1}$$

where $\boldsymbol{\theta}$ are the *log-odds* (or logit) of $\mathbf{Y} = \mathbf{1}$ and $\boldsymbol{X}$ is the $N \times K + 1$ design matrix. Since the simulation is stochastic, we replicate the design $\boldsymbol{X}$ a total of $M$ times (with each replication using different (random) pseudo-random number streams). In Li (2011), $\hat{\boldsymbol{\beta}}(r)$ is defined as the estimate of $\boldsymbol{\beta}$ from the $r-$th replication of the design $\boldsymbol{X}$ and $\mathbf{B}(M) = \frac{1}{M}\sum_{r=1}^{M}\hat{\boldsymbol{\beta}}(r)$ is the average of the estimated effect coefficients from $M$ replications. The sample covariance matrix used in the decision criteria of the test is:

$$\hat{\boldsymbol{\Sigma}}(M) = \frac{1}{M}\sum_{r=1}^{M}\frac{(\hat{\boldsymbol{\beta}}(r) - \mathbf{B}(M))(\hat{\boldsymbol{\beta}}(R) - \mathbf{B}(M))^T}{M-1}. \tag{2}$$

It can easily be shown that rank$(\hat{\boldsymbol{\Sigma}}(M)) \leq M$, so that if the number of replications is less than the number of factors then $\hat{\boldsymbol{\Sigma}}(M)$ is not of full rank and not invertible. However, in the hybrid method, we have an initial screening step where the coefficient estimates are compared to a threshold to partition into potentially important and unimportant factors. In this case, $K_{CSB}$ assumed unimportant factors are assigned to the CSB procedure and the remaining $K_{Li} = K - K_{CSB}$ factors are handled by Li (2011)'s alternative method to CSFD. Therefore, the invertibility condition requires only $M \geq K_{Li}$, but the likelihood of satisfying this condition is directly related to the true values of the coefficients and the threshold, and for some thresholds, we have $K_{Li} \approx K$. Selection of these thresholds is investigated in our numerical study in Section 3.3.

Consequently, in this paper, we instead propose to continue to use the original CSFD and CSB for binary response cases instead of those proposed by Li (2011). The asymptotic distribution of the estimator of $\boldsymbol{\theta}$, namely $\hat{\boldsymbol{\theta}}$, is:

$$\log\left(\frac{\hat{\mathbf{p}}}{1-\hat{\mathbf{p}}}\right) = \hat{\boldsymbol{\theta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \xrightarrow{d.} \text{Normal}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}) \tag{3}$$

where $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{X}(\boldsymbol{X}^T\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^T$ where $\widehat{\boldsymbol{W}} = (\omega_{ii})$ is an $N \times N$ diagonal matrix with $\omega_{ii} = \hat{p}_i(1-\hat{p}_i)$ and $\hat{p}_i = \exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})[1+\exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})]^{-1}$. The asymptotic normality of the maximum likelihood estimates of $\hat{\boldsymbol{\beta}}$ and the conditions for the asymptotic normality are detailed by Beer (2001) (refer to p. 27-31 for the conditions).

The CSFD method introduced by Shen and Wan (2009) involves a sequential hypothesis testing procedure with a fractional factorial design, and ground their approach on the observation that since $\hat{\beta}_k(r)$ (the $k-$th effect estimate from the $r-$th replication) is a linear combination of the responses for any $k, r$, $\hat{\beta}_k(r)$ is also normally distributed, and hence $\hat{\beta}_k(r)$ are identically and independently (*iid*) normally distributed. Shen and Wan (2009) then argue that "due to the *iid* normality of $\hat{\beta}_k(r)$, all the proofs of the qualification of the fully sequential testing procedures in Wan et al. (2003) and Wan and Ankenman (2006) are also valid for CSFD."

The critical point in the applicability of the CSFD technique is the normality of the model parameters, which can follow from either the normality of the data or any other asymptotic property of the estimation method

used for model fitting. As shown by Beer (2001), the maximum likelihood estimates of parameters of the logistic regression model in Eq. (1) follow an asymptotically normal distribution. Therefore, following Shen and Wan (2009), implementation of both the CSFD and CSB methods with the logistic regression model is appropriate. Note that there is no need to assume normality of log-odds ratios which Li (2011) used in their approach. Li (2011) also used 2-level fractional factorial designs of resolution III, by simply using the full factorial of $l$ factors, where $l$ satisfies $2^{l-1} \leq K \leq 2^l$ (see Montgomery (2019)). We compare below if the use of the D-optimal designs provided by Yang et al. (2011) is preferable (where $U_k$ and $V_k$ are respectively the lower and upper bounds for the $k-$th factor):

**Theorem 1.** *Under a logistic regression model with K factors, a design $\xi^*$ is a D-optimal design if $\xi^* = \{(C^*_{\ell 1}, 1/2^K)\&(C^*_{\ell 1}, 1/2^K), \ell = 1, \ldots, 2^{K-1}\}$, where $(C^*_{\ell 1})^T = (1, a_{\ell,1}, \ldots, a_{\ell,K-1}, c^*)$ and $(C^*_{\ell 2})^T = (1, a_{\ell,1}, \ldots, a_{\ell,K-1}, -c^*)$,*

$$a_{\ell,k} = \begin{cases} U_k & \text{if } \lceil \frac{\ell}{2^{K-1-k}} \rceil \text{ is odd}, \\ V_k & \text{if } \lceil \frac{\ell}{2^{K-1-k}} \rceil \text{ is even}, \end{cases}, \ \ell = 1, \ldots, 2^{K-1};\ k = 1, \ldots, K-1$$

*and $c^*$ minimizes $f(c)$, with $f(c) = c^{-2}(\Psi(c))^{-m-1}$, where $\Psi(x) = [P'(x)]^2/[P(x)(1-P(x))]$.*

Note that smaller designs (than the full factorial in Theorem 1) can be generated from selecting the required columns from a normalized Hadamard matrix and applying similar steps as described in Yang et al. (2011). For details on Hadamard matrices and their use, we refer to Georgiou et al. (2003). Here we specify $U_k$ and $V_k$ following the procedure described in Section 3.3 of Yang et al. (2011). The elements of the resulting design matrix are $+1$ and $-1$ for all but the last column, while the last column is multiplied by $c*$. This way the generated design has the same information matrix as the optimal design and thus is optimal. These are the optimal designs that were used in the following simulations. Note that, as described in Section 3.3 of Yang et al. (2011), the initial design was obtained by selecting columns of a suitable Hadamard matrix to generate a design having the minimum number of runs possible.

## 3 MONTE CARLO EXPERIMENTS

In this section we report on simulation experiments conducted to investigate possible sensitivities of the hybrid method to user defined parameters (see below). We also compare the effectiveness of the D-optimal design to the more common fractional factorial design, both having the same minimum number of runs.

### 3.1 Experiment setup

The experimental setup requires choosing values for the number of factors under investigation; the number of active factors in the simulations; the magnitudes of the coefficients of the active and inactive factors; the desirable Type I error and power of the method; the thresholds defining unimportant ($\Delta_0$) and important ($\Delta_1$) factors; the initial factor assignment threshold ($\Delta$); the initial number of replications for CSFD and CSB; and the number of pre-screening replications.

The screening aims to have no more than $\alpha\%$ of the inactive factors (with coefficients less than $\Delta_0$) to be misclassified as active (where $\alpha$ is the typical confidence level), and no more than $(1 - power)\%$ of the active factors (with coefficients more than $\Delta_1$) to be misclassified as inactive. So, $\Delta_0$ and $\Delta_1$ directly affect the Type I and Type II error rates of the screening process respectively. Typically, there is a preference in controlling the Type II errors during screening, as it is better to include (erroneously) non-active factors in subsequent meta-model construction than to exclude (prematurely) active factors during screening. For further details refer to the flow charts given in Figure 1 and the references listed there.

Specifically, we investigate the following default values: $100$ factors with $10$ ($10\%$) being significant; coefficients of active factors set to $4.0$; significance level set to $5\%$; with complementary power of $95\%$; $\Delta_0$ and $\Delta_1$ set to $1$ and $3$ respectively; the factor assignment threshold set to $3$; and the initial number of replications and pre-screening replications set to $10$.

### 3.2 D-Optimal vs. fractional factorial design

Figure 2 displays the performance of the two designs as well as the sensitivity of both to the number of pre-screening replications. Even though both return Type I and Type II errors within the desirable bounds $(0.05)$, there are some differences worth mentioning. Importantly, for most values of the number of pre-screening replications, we observe that the errors from using the D-optimal design are smaller than those of the fractional

factorial design. It consistently achieves smaller Type I errors than the fractional factorial design does, and only for large numbers of pre-screening replications do the two designs converge in performance. For Type II errors, we observe that the D-optimal design is only less effective in the case of very small number of pre-screening replications (where in contrast, its superiority regarding Type I errors is greatest). When the number of pre-screening replications is larger, the D-optimal design become preferable regarding Type II errors.

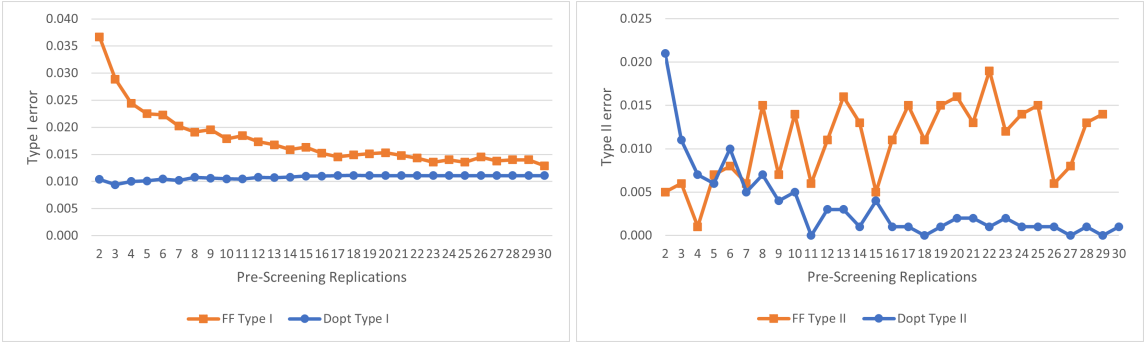


**Figure 2**. Comparison of Fractional Factorial design (FF) with D-optimal design (Dopt) for binary responses

### 3.3 Factor assignment threshold

The next parameter studied was the factor assignment threshold ($\Delta$) as this plays a crucial role in the efficiency of the screening method. We kept all other parameters fixed at their default values and varied the factor assignment threshold $\Delta$ from 1 to 11. The design used was the D-optimal design due to its performance noted above. The percentage of factors assigned to the CSFD method is displayed in Figure 3 (the number of pre-screening replications is set to either 2 or 20) and the resulting Type I and Type II errors are presented in Figure 4.

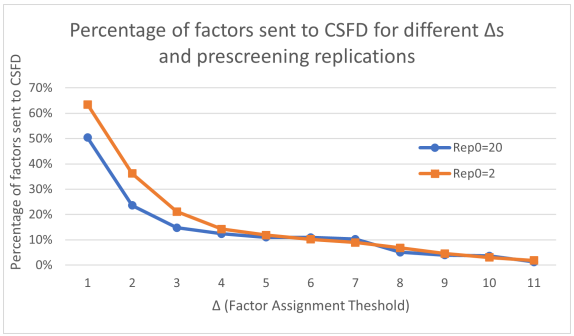


**Figure 3**. Partitioning of factors for different values of the factor assignment threshold

The number of factors assigned to CSFD decreases rapidly as $\Delta$ increases, but is not particularly sensitive to the number of pre-screening replications. From Figure 4, both Type I and II errors are low for threshold values up to 6, however for larger thresholds the Type II error explodes. This means that we will only be able to identify a small percentage of the active factors if we set a value for the threshold that significantly overestimates the true coefficients magnitudes. Any threshold lower than 6 is appropriate but optimal performance was achieved for values between 3 and 5.

### 3.4 Unimportant and important thresholds

The simulation setup to study $\Delta_0$ and $\Delta_1$ is as follows: the insignificant factors' coefficients are set to 0.41; the factor assignment threshold is set to 5; and the remainder are set to their default values.The results are

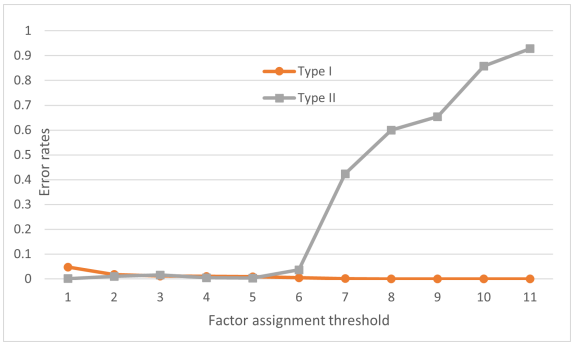


**Figure 4**. Error sensitivity to different values of the factor assignment threshold

presented in Figure 5.

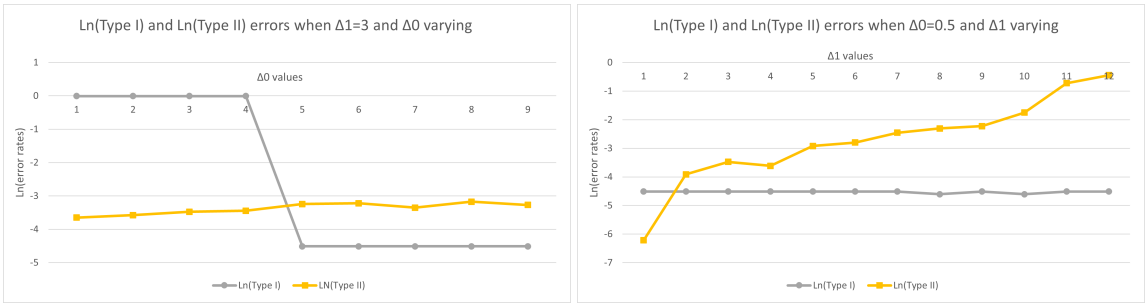


**Figure 5**. Sensitivity to different values of $\Delta_0$ (left) and $\Delta_1$ (right)

The left image of Figure 5 displays the (log of the) Type I and Type II errors when $\Delta_0$ varies from $0.1$ to $0.9$ while $\Delta_1 = 3$. It is clear that when $\Delta_0 < 0.41$ (the coefficient value of the inactive factors) the method does not control the Type I error and almost all inactive factors are misclassified as active. When $\Delta_0 > 0.41$, we note that the assumptions of the method are satisfied and Type I error is controlled as desired. Type II error is consistently lower that the $(1 - power) = 5\%$, as expected, since $\Delta_1$ is smaller than the coefficient of the active factors.

Typically, the coefficients of the inactive factors are expected to be zero, but these simulations indicate that if, due to errors for example, some inactive coefficients become nonzero then the method is very sensitive to the choice of $\Delta_0$. So, in such cases there will be large Type I errors when $\Delta_0$ is set incorrectly by the user and specified in such a way to be less than the estimated coefficients of the inactive factors.

The right image of Figure 5 displays the Type I and Type II errors when $\Delta_1$ varies from $1$ to $12$ while $\Delta_0 = 0.5$. It is clear that when $\Delta_1$ is more than $4.0$ (the minimum coefficient of the active factors) the method tries but cannot control the Type II error and some active factors are misclassified as inactive. When $\Delta_1 < 4.0$, we again note that the assumptions of the method are satisfied and Type II error is controlled as desired. Type I error is consistently lower than $5\%$, as expected, since $\Delta_0$ is correctly set higher than the maximum coefficient of the inactive factors. Incorrectly specifying $\Delta_1$ will result in not identifying some of the significant factors, which will inflate Type II errors and is particularly undesirable during the screening stage of simulation analytics.

## 4 DISCUSSION

To identify the important input factors driving the binary response of stochastic simulations (e.g., mission success of combat simulations) an effective screening procedure is required. Somewhat surprisingly, the only published resource found by the authors specific to this problem is that of Li (2011), who modified elements of the hybrid screening approach of Shen et al. (2010) for linear meta-models. However, the requirement of the modified approach in using the sample covariance matrix of the regression estimates was shown in this paper

to be susceptible to singularity issues due to rank deficiency. In contrast, we claim that the original Shen et al. (2010) remains a valid screening approach, based on asymptotic properties of the estimation method.

Numerical experiments found that this hybrid method can work well when the required parameters are specified correctly. It worked efficiently with both the typical fractional factorial design and the alternative D-optimal design specifically for binary responses. As expected the D-optimal design outperforms the fractional factorial design, but the differences were relatively minor and the method seemed to behave well under both designs.

The correct specification of the user parameters is related to the problem being studied, but is influenced by parameters not under the researcher's control, for example the correct assumption of the fitted model and the identification of the variability of the true active/inactive coefficients. Such parameters are $\Delta_0, \Delta_1$ and the factor assignment threshold $\Delta$. The method was not found to be robust to misspecification of these user parameters and the results in such cases might be misleading at best. This disadvantage makes its use highly problematic in practical situations where prior knowledge of the true underling model and range of significant/insignificant coefficients might be limited.

Further research is thus required in identifying a robust factor screening method for binary responses from stochastic simulations. Similarly, for response variables that are discrete but bounded, such as the number of combat assets lost or remaining, an effective screening approach remains elusive. Such generalisations are not straightforward and more research is needed in these directions.

## ACKNOWLEDGEMENT

The Commonwealth of Australia (represented by the Defence Science and Technology Group) supported this research through a Defence Science Partnership agreement.

## REFERENCES

Beer, M. (2001). Asymptotic properties of the maximum likelihood estimator in dichotomous logistic regression models. Master's thesis, University of Fribourg, Switzerland. Diploma Thesis submitted to the University of Fribourg, Switzerland.

Fedorov, V. (1972). *Theory of Optimal Experiments*. Probability and Mathematical Statistics. Academic Press.

Georgiou, S., C. Koukouvinos, and J. Seberry (2003). *Hadamard matrices, orthogonal designs and construction algorithms*. Designs 2002: Further computational and constructive design theory. Springer.

Gill, A., D. Grieger, M. Wong, and W. Chau (2018). Combat simulation analytics: Regression analysis, multiple comparisons and ranking sensitivity. In *Proceedings of the 2018 Winter Simulation Conference*, WSC '18, pp. 3789–3800. IEEE Press.

Li, M. (2011). Simulation factor screen in binary response models. Master's thesis, West Virginia University. Graduate Theses, Dissertations, and Problem Reports. 2214.

Montgomery, D. C. (2019). *Design and Analysis of Experiments, 10th Edition*. General & Introductory Industrial Engineering. Wiley.

Shen, H. and H. Wan (2009). Controlled sequential factorial design for simulation factor screening. *European Journal of Operational Research 198*(2), 511–519.

Shen, H., H. Wan, and S. M. Sanchez (2010). A hybrid method for simulation factor screening. *Naval Research Logistics 57*(1), 45–57.

Wan, H. and B. Ankenman (2006). Two-stage controlled fractional factorial screening for simulation experiments. *Journal of Quality Technology 39*(2), 126–139.

Wan, H., B. Ankenman, and B. Nelson (2003). Controlled sequential bifurcation: A new factor-screening method for discrete-event simulation. In *Proceedings of the 2003 Winter Simulation Conference*, Piscataway, NJ, pp. 565–573. Institute of Electrical and Electronics Engineers.

Yang, M., B. Zhang, and S. Huang (2011). Optimal designs for generalized linear models with multiple design variables. *Statistica Sinica 21*(3), 1415–1430.