2013

Efficient, nearly orthogonal-and-balanced, mixed designs: an effective way to conduct trade-off analyses via simulation

Journal of Simulation (2013) 7, 264–275

www.palgrave-journals.com/jos/

Efficient, nearly orthogonal-and-balanced, mixed designs: an effective way to conduct trade-off analyses via simulation

H Vieira Jr[1], SM Sanchez[2*], KH Kienitz[1] and MCN Belderrain[1]

[1]*Technological Institute of Aeronautics, São Jose dos Campos, Brazil; and* [2]*Operations Research Department, Naval Postgraduate School, Monterey, USA*

Designed experiments are powerful methodologies for gaining insights into the behaviour of complex simulation models. In recent years, many new designs have been created to address the large number of factors and complex response surfaces that often arise in simulation studies, but handling discrete-valued or qualitative factors remains problematic. We propose a framework for generating a design, of specified size, that is nearly orthogonal and nearly balanced for any mix of factor types (categorical, numerical discrete, and numerical continuous) and mix of factor levels. These new designs allow decision makers structured methods for trade-off analyses in situations that are not necessarily amenable to other methods for choosing alternatives, such as simulation optimization or ranking and selection approaches. These new designs also compare well to existing approaches for constructing custom designs for smaller experiments, and may also be of interest for exploring computer models in domains where fewer factors are involved.
Journal of Simulation (2013) **7,** 264–275. doi:10.1057/jos.2013.14

Keywords: simulation; design of experiments; integer programming; statistics; defense studies

1. Introduction

Simulation is often used to investigate alternatives. Before setting up a new manufacturing plant, a company might be interested in exploring different types of production and material handling equipment, facility layouts, workforce capabilities, buffer sizes, and processing protocols to determine suitable choices. Similarly, military decision makers might wish to explore the potential impact of different tactics, equipment, training, and logistical support policies on operational efficacy. Alternatives are instantiated as different combinations of settings for model factors (ie, inputs, parameters, or components) that reflect key characteristics of the real-world situation.

There are many methods for trying to identify the so-called 'optimal' or 'good' alternatives. A broad categorization includes simulation optimization, Ranking and Selection (R&S) procedures, and designed experiments. Simulation optimization methods seek to identify a combination of factor settings that yields the best outcome. R&S approaches examine a finite (often, relatively small) set of alternatives and seek to select either the best one subject to some minimum practical difference or a subset of random size that contains the true best; all R&S procedures include some sort of guarantee on the probability of correct selection. Experimental design specifies the configurations that will be examined, and the choice of the design has bearing on the type of analyses that can subsequently be conducted.

*Correspondence: SM Sanchez, Operations Research Department, Naval Postgraduate School, 1411 Cunningham Rd, Monterey, CA, 93943, USA.
E-mail: ssanchez@nps.edu

We assert that neither simulation optimization nor R&S is directly appropriate for those interested in performing trade-off analyses for complex scenarios. Simulation optimization methods end up yielding a single alternative that may be a local (not global) optimum. They do not provide information about so-called 'knees in the curve', where increasing one type of resource beyond a certain point leads to either diminishing or increasing returns on overall performance. Similarly, simulation optimization methods may not reveal the locations of tipping points, where small changes in the input result in fundamentally different output behaviour.

The same is true of indifference-zone R&S methods, where a single alternative is designated as 'best' at the end of experimentation. Subset selection methods can be used for first getting a set of good alternatives with respect to one performance measure, and then for making the final selection based on different criteria. But since these procedures are still based on the notion of the 'best' (minimum or a maximum) primary performance measure, they do not directly apply when the decision maker cares about intermediate response values.

To further complicate matters, decision makers are not necessarily interested in the 'best' option with regard to any single criterion at the end of the simulation study. Instead, the overall decision is often made by assessing several performance measures. Consequently, decision makers need methods for gaining a broad understanding of the simulation's behaviour in order to make effective trade-offs. Some of these measures (eg, cost, throughput, or implementation time) might be quantitative, but others (eg, ease of implementation, subjective assessments of future risk) might be qualitative.

For these reasons, well-designed experiments are the most suitable approach if one wants to explore trade-offs. Fisher (1925) pioneered the design of experiments (DOE) field, where the basic principles are the use of *randomization*, *replication*, and *control* to allow the analyst to make statistically valid inferences about the behaviour of a system. As noted by Montgomery (2005, p 21), 'there is not a single area of science and engineering that has not successfully employed statistically designed experiments'. Simulation is one area that has benefitted; see, for example, Santner *et al* (2003), Kleijnen (2007, 2008), Law (2007), or Sanchez and Wan (2012) for a general discussion of simulation experiments. Factory design (Montevechi *et al*, 2010), large-scale networks (Van Vorst *et al*, 2012), and defence applications (Sanchez *et al*, 2012) are just a few of the many application areas. However, the basic designs in standard DOE texts often do not suffice. The complexity of many simulation models—in terms of both the number of model inputs (or factors) that can be explored, and the complexity of a multidimensional response surface—means that designs intended for use in physical experiments cannot be used in the simulation environment without making restrictive or unwarranted assumptions. Analysts need an expanded portfolio of designs in order to effectively and efficiently conduct large-scale simulation experiments (Kleijnen *et al*, 2005). The designs we propose here are part of this expanded portfolio.

The title of this paper mentions several desirable design properties that we now define.

- *Mixed designs* are those capable of handling different factor types (categorical, discrete, and continuous) and/or discrete factors with different numbers of levels (eg, Factor 1 with 10 levels, Factor 2 with 5 levels etc).
- A design is *balanced* if, for every column, every factor level occurs equally often. We call a design *nearly balanced* if the ratio of the actual to ideal number of occurrences is sufficiently close to one.
- A design where the maximum absolute pairwise correlation between any 2 quantitative factors is 0 is said to be an *orthogonal design*. If this maximum absolute pairwise correlation is sufficiently small ($\leqslant 0.05$), this is considered a *nearly orthogonal design*.
- Finally, we characterize a design as *efficient* if the number of design points is acceptable. This concept is subjective and it is problem driven.

The above concepts are important for several reasons. Simulation problems usually have different factor types and factor levels, and designs that accommodate this are needed. The balance property allows correct analysis of heteroscedastic non-normal experiments (Bathke, 2004). Orthogonality makes it possible to model the effect of one factor independently of other factors (see, eg, Montgomery (2005, p 91) and Ryan (1997, p 122)). Finally, despite the ready availability of high-speed computing processors, brute-force computation cannot be used to explore large-scale simulation experiments. Real-world simulation studies face restrictions due to time, cost, number of computers available for experimentation and so on. They need efficient designs, although the number of design points is not the overriding consideration.

An implicit characteristic of a design for trade-off analyses is its space-filling behaviour. When factors are continuous, space-filling designs are useful for exploratory studies because they provide insight about the simulation behaviour throughout the region of interest. An analogy for discrete-valued factors is that they take on many (perhaps all) of the potential levels of interest. For example, a design where x assumes levels $\ell_x \in \{0, 1\}$ (in weeks) is less space-filling than a design where x assumes levels $\ell_x \in \{1,2, \ldots, 7\}$ (in days). For categorical factors, we assume that ℓ_x may need to be large in order to adequately reflect the complexity of the real-world situation being modelled and/or that a moderate to large number of categorical factors exists.

There are several ways to assess the quality of designs. For traditional designs, the most usual methods are known as alphabetic optimality criteria and include, but are not restricted to, the *A*-, *D*-, and *I*-optimality, where the most widely known and cited in the literature is *D*-optimality. We refer the readers to Atkinson and Donev (1992) for further information about these alphabetic criteria, but note that despite the term 'optimality' in the names these designs are often generated by heuristics. Various 'optimized' Latin hypercubes (LHs) are also used for analysing simulation experiments. These are constructed optimizing a criterion, though once again the methods are often heuristic and a true optimal solution is not guaranteed. Space-filling measures are commonly used to construct and evaluate these designs. Maximin LHs maximize the minimum distance between two points, while minimax LHs minimize the maximum distance between two points. Santner *et al* (2003) have further information about these optimized LHs. Others, including Cioppa and Lucas (2007), Joseph and Hung (2008), and Hernandez *et al* (2012), have explored ways to construct designs that perform well in terms of orthogonality and space-filling behaviour. Lu *et al* (2013) propose an algorithm to construct small-sample designs that perform well on three criteria, by identifying designs that lie on the Pareto frontier (ie, are not dominated in all criteria by any other design). As our research is focused on correctly identifying the factors that most drive the outcome, rather than seeking models that make accurate predictions, we are more interested in orthogonality (or near-orthogonality) than these other criteria. Nonetheless, we will show that our designs perform well on a variety of measures.

This paper focuses on design, but design and analysis go hand in hand. We have found that metamodel construction and partition trees (also known as classification and regression trees) are particularly useful, but other graphical and analytic techniques are possible. Kriging or stochastic kriging metamodels make few assumptions about the response behaviour. A detailed discussion of potential analysis approaches is beyond the scope of this paper, but the reader should be aware that the choice of the design places restrictions on the types of analyses that can subsequently be conducted. For example, two-level designs such as fractional factorials or Box-Behnken (BBH) designs do not allow for the

estimation of quadratic effects. Some two-level designs, such as saturated factorials, do not allow for the estimation of any interaction effects.

Because we wish to facilitate trade-off analysis, we seek space-filling designs; they can be used to fit (very) high order metamodels to the data if the simulation behaviour is complex, and they are amenable to trade-off analysis using non-parametric techniques such as partition trees. At the same time, our designs should still have very good properties if the true underlying response behaviour is simple. For example, we want good estimates of the factor effects if the I/O relationship for a particular response of interest is characterized well by a linear metamodel involving a few inputs. This means that orthogonality (or near-orthogonality) is also a desirable property. Optimization methods can be applied directly to the metamodels to identify (potentially) good alternatives, but, as we discuss above, better results may be achieved when decision makers gain a broad understanding of the simulation's behaviour, rather than receive a specific recommendation. Robust design is often of interest for large-scale simulation studies, because decision makers may be interested in identifying combinations of the controllable decision factors that lead to good results across a variety of combinations of noise factors; see, for example, Kleijnen *et al* (2005), Dellino *et al* (2012), or Sanchez and Wan (2012). In this context, there is often less of a concern about the ability to identify noise factor contributions; even so, designs capable of handling large numbers of factors are necessary. The analysis flexibility provided by space-filling designs makes them particularly useful when there is little prior knowledge about the underlying responses.

2. Notation

Let M denote an $n \times K$ design matrix, with elements m_{rc}, and for notational convenience let $\bar{c}$ and s_c denote the mean and standard deviation of column c, respectively. Here, n is the number of design points and K is the number of factors. In the statistical literature, n is often referred to as the number of runs, but we use design points to avoid confusing 'run' with the act of running the simulation model. For stochastic simulations, the design may be replicated multiple times.

The entries in column x of M are the values 1, 2, …, ℓ_x, so ℓ_x represents the number of distinct levels of factor x. An expanded matrix $\tilde{\mathrm{M}}$ will also be useful for presenting our Mixed Integer Programming (MIP) formulation. Here, a single column is used for any quantitative factor (continuous or discrete). A qualitative factor x with ℓ_x levels in M needs $\ell_x - 1$ indicator variable columns in $\tilde{\mathrm{M}}$, denoted by $x^1, x^2, \ldots, x^{\ell_x - 1}$. The entry in row r for the indicator variable column x^i has the form:

$$x_r^i = \begin{cases} 1 & \text{if } x_r = i,\ i < \ell_x; \\ -1 & \text{if } x_r = \ell_x; \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Other indicator variable codings are possible, such as a two-level 0/1 coding with the omitted factor representing the baseline, but this three-level coding assures that when regression models are fit to the resulting data, the intercept represents the overall mean response. Note that for quantitative factors, the levels for factor x in M (or $\tilde{\mathrm{M}}$) can be converted from the coded levels 1, 2, …, ℓ_x to the natural units of the problem by simple scaling before running the experiment. Similarly, for qualitative factors, the coded entries in M can be converted to appropriate labels.

Let ρ_{map} denote the maximum absolute pairwise correlation between any two columns of $\tilde{\mathrm{M}}$. Let $w_{i,x}$ denote the number of occurrences of level i for factor x in M. Let δ_x denote the imbalance associated with column x of M, that is,

$$\delta_x = \max_{i=1, \ldots, \ell_x} \left| \frac{w_{i,x} - (n/\ell_x)}{(n/\ell_x)} \right|. \quad (2)$$

Let $\delta \equiv \max\{\delta_x, x = 1, \ldots, K\}$ denote the maximum imbalance of design M.

3. Test cases

We are interested in designs that can handle mixed factor types. As a test case, let L^1 denote one set of 10 factors that includes a single factor with each of 2, 3, …, 11 distinct levels. L^i will denote a larger design problem with $10i$ factors, that is, i factors with each of 2, 3, …, 11 levels. Let S^i denote the set of $15i$ factors that augments L^i with $5i$ continuous-valued factors. We will at times use L^i and S^i as test cases when comparing different design construction methods. Design characteristics such as the number of design points n, the non-orthogonality ρ_{map}, and the design imbalance δ will be important in these comparisons.

4. Modifying designs constructed for categorical factors

Orthogonal arrays (OAs) have played an important role in experimental design (see Hedayat *et al*, 1999 for more information). Consider an $n \times K$ design matrix M. If any subarray of size $n \times g$ contains all possible combinations of values equally often as rows, then M is said to be an OA of '*strength g*'. OAs of strength 2 can be used to estimate main effects models for categorical factors, those of strength 3 can be used to estimate main effects and two-way interactions, and so on. The strength corresponds to special types of balance, and means that OAs can be used for exploring quantitative and categorical factors. However, this flexibility comes at the cost of large designs: $\ell_x - 1$ degrees of freedom are needed to estimate the main effects for a categorial factor with ℓ_x levels and $(\ell_x - 1)(\ell_y - 1)$ degrees of freedom are needed to estimate the two-way interaction effects between two categorical factors x and y. This means that $n \geqslant 1 + \sum_{x=1}^{K} (\ell_x - 1)$ design points are needed for a strength 2 experiment involving K categorical factors, and at least $\sum_{x=1}^{K} \sum_{y=1}^{K} (\ell_x - 1)(\ell_y - 1)$ additional design points are required in order to estimate all two-way interactions.

Now suppose that the experiment includes quantitative factors and categorical factors. If OAs are to be used, a (numerical) discrete factor x with ℓ_x levels will use $\ell_x - 1$ degrees of freedom

as above. In contrast, if x is treated as a quantitative factor, then a single degree of freedom is sufficient for estimating the main effect of x (two degrees of freedom can be used to estimate a quadratic relationship and so forth). Clearly, treating the factor as quantitative is more efficient if a parsimonious representation of the response's dependence on x can be obtained.

OAs are most efficient if all the ℓ_x are small, so there is a temptation to set $\ell_x = 2$ for any quantitative factor x. However, the resulting designs will have poor space-filling behaviour, and so are far less useful for trade-off studies than other designs. But if the ℓ_x are large, then the size of the OA can be immense. The number of design points must be divisible by the lowest common multiple (LCM) of the numbers of factor levels to achieve perfect balance. In real-word experiments, the LCM usually is a big number: the LCM for L^1 is 27 720.

Another difficulty in using OAs for trade-off analysis is the availability of suitable OAs. Extensive online libraries are available in Sloane (2007) and Kuhfeld (2010), but these designs have largely been developed for other types of applications, and tend to involve relatively small numbers of factors and/or limited mixes of the ℓ_x. Crossing several small OAs to obtain a larger OA can result in extremely large designs. For example, in order to obtain a design that could be used for a L^1 by crossing OAs from Sloane's online library, the OA would require 9 979 200 design points. This is only 25% of the amount required by a full factorial, but nonetheless it is an enormous number. A corresponding design capable of exploring L^{20} would require over 2.3×10^{33} design points, which is effectively infinite.

Some work has been done on nearly orthogonal (rather than completely orthogonal) arrays. These typically require software algorithms to generate solutions, and the solutions may depend on the pseudo-random numbers used within the algorithms, but once again they are not intended for very large-scale simulation studies. For example, the Gendex software developed by Nguyen (http:// designcomputing.net/gendex/noa/) limits the user to specifying designs with at most 100 design points, and factors with at most 17 (categorical) levels. L^1 can be explored using 56 design points: 50 runs of the programme, each choosing the best design based on 100 trials, yielded designs with $\delta = 0.1786$ and $0.10 \leqslant \rho_{map} \leqslant 0.46$ when they were applied to discrete factors. These designs do not meet our near-orthogonality criteria of $\rho_{map} \leqslant 0.05$. L^1 can also be explored in 56 design points using the Custom Design capability in JMP 9.0 Professional: 25 runs of the programme, each choosing the best design based on 100 trials, yielded designs with $0.214 \leqslant \delta \leqslant 0.375$ and $0.137 \leqslant \rho_{map} \leqslant 0.246$. Once again, these designs do not qualify as nearly orthogonal. The minimum number of design points required rises as the number of blocks increases (eg, L^{10} can also be analysed in JMP 9 using 551 design points with $\rho_{map} = 0.0878$ and $\delta = 0.1325$), but the memory requirements rapidly become prohibitive once these designs get larger. Note that the algorithms used by both Gendex and JMP involve randomness, so choosing a 'good' design involves invoking the design construction methods a number of times with different random number seeds.

The lowest achievable design matrix imbalance is not monotonic in n. Figure 1 illustrates the analytic minimum design imbalance for L^1, which is identical to that of L^i or S^i $(i > 1)$. We need $n \geqslant 46$ to ensure $\delta \leqslant 0.20$, $n \geqslant 100$ to ensure $\delta \leqslant 0.10$, and $n \geqslant 200$ to ensure $\delta < 0.05$; even with $n = 500$, the minimum imbalance is 0.011. This value may not be achievable if ρ_{map} is constrained.

We will revisit nearly orthogonal arrays (NOAs) for small-scale experiments later in this paper, and compare them with the new designs we propose. For now, note that OAs and NOAs are not readily available for mixed factor experiments involving several categorical, discrete, or continuous factors. If they do exist, they will likely require an excessively large number of design points.

5. Modifying continuous-factor designs for use with discrete factors

An alternative to adapting OAs (or other categorical factor designs) for use with discrete factors is to adapt designs originally constructed for quantitative, continuous-valued factors.

LHs are probably the most familiar space-filling designs. Randomly generated LHs have been widely used for computational experiments (Sacks *et al,* 1989). They tend to have good space-filling and orthogonality behaviour if $n \gg K$, but when $n \approx K$ they can perform quite poorly. Cioppa and Lucas (2007) constructed efficient, space-filling, nearly orthogonal Latin hypercubes (NOLHs) that have proven useful for investigating continuous factors in a number of studies. To overcome the limited combinations of K and n for which NOLHs were available, Hernandez *et al* (2012) developed a MIP approach that allows for the construction of NOLHs for non-saturated cases $(2 < K < n)$, and reviewed other NOLH designs available in the literature.

One issue relating to all of the LH designs is that they were initially intended for continuous-valued factors. Applying them to discrete-valued factors requires rounding. A limited amount of rounding is acceptable, but if there are several factors with small numbers of levels, coupled with a relative small number of design points, this can destroy the near-orthogonality of the designs. Figure 2 illustrates this phenomenon for two different size problems: 15 factors in 18 design points and 135 factors in 162

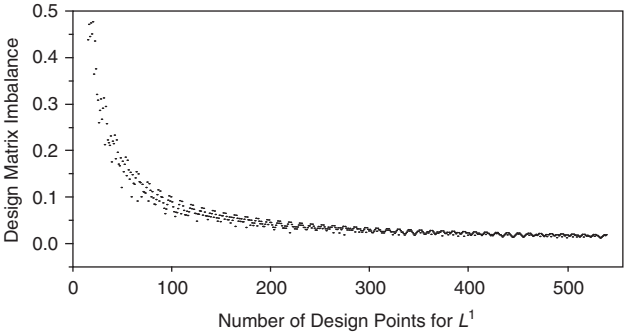

Figure 1 Minimum design matrix imbalance, as a function of n, for L^1.

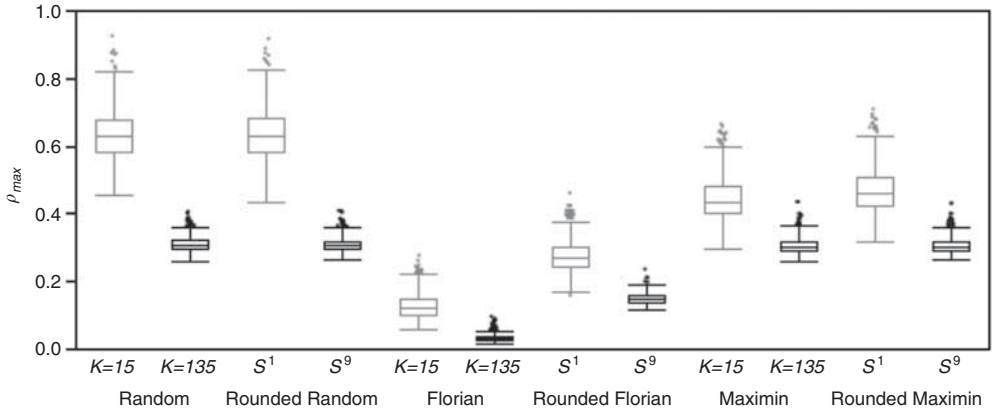

Figure 2 Distributions of ρ_{map} for various designs involving 15 factors in 18 design points (grey), and 135 factors in 162 design points (black). Each box plot is based on 1000 replications.

design points. The first pair of box plots show the distribution of ρ_{map} based on 1000 randomly generated LH designs. Note that these suffer from a lack of orthogonality: ρ_{map} is substantially higher than 0.05 for all cases. The second pair of box plots indicates that modifying these randomly generated LH designs to obtain the desired levels for the limited-level factors in S^1, and S^9 has little impact on the respective ρ_{map} values. The third pair of box plots shows that ρ_{map} drops dramatically when the method of Florian (1992) is applied to the initial, randomly generated, continuous-valued designs. This approach yields designs that meet our near-orthogonality criterion of $\rho_{map} <$ 0.05 roughly 96% of the time when $K = 135$; the minimum ρ_{map} was 0.057 when $K = 15$. However, as the fourth pair of box plots shows, this improvement in ρ_{map} is not maintained when the columns in these designs are, in turn, rounded for S^1 and S^9. The last two pairs of box plots show that rounding has little impact on ρ_{map} for maximin LH designs, but these do not meet our near-orthogonality criteria. A nearly orthogonal design for continuous factors can deteriorate substantially when it is modified for discrete-valued factors with limited numbers of levels. In fact, it appears that difficult-to-find, nearly orthogonal designs are more likely to suffer when the factor levels are rounded. Similar problems arise when starting with other types of space-filling designs, such as uniform designs or sphere-packing designs.

If rounding a particular design M causes problems, there are a few steps the analyst can take to mitigate these problems. First, the analyst could construct a new design based on $n' > n$ design points to determine whether the reduction in granularity of the base design reduces the correlations induced by rounding. Even so, achieving good orthogonality in the presence of rounding is not guaranteed. Alternatively, the analyst could construct several designs and stack them until suitable near-orthogonality is achieved. However, this is an *ad hoc* method. If the original NOLH (for continuous factors) has n design points, then each stack has $\approx n$ design points as well. In addition, while the non-orthogonality problems associated with rounding may be overcome by constructing substantially larger designs, there are times when computing budget or time constraints make this brute force approach prohibitive.

Even if the rounding problem is solved, the resulting designs deal only with numerical factors. Although someone unfamiliar with experimental design might be tempted to use the entries in a column constructed for a discrete-valued factors as codes for levels of a qualitative factor, this can lead to severe problems in the analysis phase (see, eg, Vieira *et al*, 2011b, for details).

6. Crossing separately constructed designs

If suitable designs can be created for each type of factor separately, then these smaller designs can be crossed to obtain one that, overall, is close to orthogonal. For example, an OA design M_1 can be used for factors that are categorical or discrete with a limited number of levels. A space-filling design M_2 can be used for continuous factors, and for discrete factors with many levels of interest. However, if designs M_1 and M_2 have n_1 and n_2 design points, respectively, then the crossed design $M_1 \times M_2$ will have $n_1 \times n_2$ design points. It is difficult to obtain efficient designs when there are moderate or large numbers of factors in each category.

7. Directly constructing nearly orthogonal-and-balanced (NOAB) mixed designs

Our proposal takes a more direct approach for constructing NOAB designs for mixed factors. This builds on the work of Hernandez *et al* (2012), who proposed a MIP formulation in order to construct nearly orthogonal designs for continuous factors. The MIP was extended and enhanced in Vieira *et al* (2011a) in order to construct orthogonal designs, or improve existing OAs, for experiments involving quantitative factors with limited numbers of levels of interest. However, it is not possible to obtain orthogonal designs for every n for a specific set of factors and levels. In subsequent studies involving the MIP of Vieira *et al* (2011a), we noticed that we were often able to find

Table 1 Inputs and balance feasibility test for NOAB construction

INPUTS:
Study characteristics:
- K — Number of factors of interest
- $C(x_j)$ — Type of the jth factor x_j (1 for continuous, 2 for discrete, 3 for categorical)
- ℓ_{x_j} — Number of levels ($\leqslant n$) associated with factor x_j

Design characteristics:
- n — Number of design points ($n \geqslant K + \sum(\ell_{x_j} - 2)$, where the summation is taken over all x_j satisfying $C(x_j) \neq 1$)
- α^* — Maximum allowable ρ_{map} ($0 \leqslant \alpha^* < 1$)
- δ^* — Maximum allowable imbalance ($0 \leqslant \delta^* < 1$)

MIP and heuristic parameters:
- $t_{\min}$ — Minimum allowable time for MIP solution search
- $t_{\max}$ — Maximum allowable time for MIP solution search
- h^* — Maximum number of iterations per column
- b^* — Maximum number of macro-iterations

FEASIBILITY TEST
$\delta \leftarrow \infty$
for $j = 1, j < K$ **do**
  if $C(x_j) \in \{2, 3\}$ **do**
    $\delta_{x_j} \leftarrow \left(\frac{\ell_{x_j}}{n}\right) \max\left(\left(\left\lceil \frac{n}{\ell_{x_j}} \right\rceil - \frac{n}{\ell_{x_j}}\right), \left(\frac{n}{\ell_{x_j}} - \left\lfloor \frac{n}{\ell_{x_j}} \right\rfloor\right)\right)$
    $\delta \leftarrow \min(\delta, \delta_{x_j})$
  end
end
if $\delta > \delta^*$
  RETURN "No feasible solution exists with current balance constraints. Increase n until the feasibility check is passed, or set $\delta^* = \delta$"
else **RETURN** "Initial balance feasibility check passed"

Table 2 NOAB construction method

$b \leftarrow 0$
if $\{b < b^*\}$
  $\mathrm{M}_0 \leftarrow \emptyset$, $\tilde{\mathrm{M}}_0 \leftarrow \emptyset$
  $j \leftarrow 0$
  if $\{j < K\}$ **do**
    solution $_{j+1} \leftarrow$ "FALSE"
    $h \leftarrow 1$
    if $\{h < h^*$ **AND** solution $_{j+1}$ = "FALSE"$\}$ **do**
      $t \leftarrow t_{\min}$
      $x \leftarrow$ an $n \times 1$ vector, randomly generated from $x \in \tilde{\mathcal{B}}(n, \varsigma_{j+1})$
      if $\{ t \leqslant t_{\max}$ **AND** solution $_{j+1}$ = "FALSE" $\}$ **do**
        call MIP using $\tilde{\mathrm{M}}_j$, δ^*, t, x, ℓ_x, and $C(x)$.
        $v^* \leftarrow$ MIP objective function value
        $x^* \leftarrow$ MIP modified column vector
        $s_{x^*} \leftarrow$ standard deviation of x^*
        if $v^* \leqslant \alpha^* s_{x^*}$ **do**
          solution $_{j+1} \leftarrow$ "TRUE"
        end
        else if $\{v^* > \alpha^* s_{x^*}$ **AND** $t < t_{\max}\}$ **do**
          $t \leftarrow t + t_{\min}$
        end
        else do
          $h \leftarrow h + 1$
          $t \leftarrow t_{\min}$
        end
      end
    end
    if {solution $_{j+1}$ = "TRUE"} **do**
      if $C(x) = 3$ (ie, x is categorical) **do**
        $x^{*i} \leftarrow i^{th}$ indicator vector associated with x^* ($i = 1, \ldots, \ell_x - 1$)
        $\tilde{\mathrm{M}}_{j+1} \leftarrow \left[\tilde{\mathrm{M}}_j \quad x^{*1} \quad x^{*2} \quad \cdots \quad x^{*(\ell_x - 1)}\right]$
      end
      else do
        $\tilde{\mathrm{M}}_{j+1} \leftarrow \left[\tilde{\mathrm{M}}_j \quad x^*\right]$
      end
      $\mathrm{M}_{j+1} \leftarrow [\mathrm{M}_j \quad x^*]$
      $j \leftarrow j + 1$
    end
  end
  if {solution $_K$ = "TRUE"} **RETURN** M_K
  else $b \leftarrow b + 1$
end
RETURN "No solution found that meets near-orthogonality criteria"

new columns with very low maximum absolute pairwise correlation but were not able to find orthogonal ones. We also found that the balance constraints were often tight. In other words, even though the use of discrete (*versus* categorical) factors reduces the minimum number of design points required, trying to simultaneously achieve perfect orthogonality and balance is hard.

We create NOAB designs one column at a time, starting with a randomly generated column that satisfies the balance constraint. We then use one of three potential MIP formulations to add the next column, where the formulation depends on the type of factor (discrete, continuous, or categorical). Continuous factors are sampled at n distinct, equally spaced levels, so the columns for the continuous factors constitute a lattice LH. Details appear in the Appendix, but a brief overview of our approach follows. Algorithmic descriptions of the initialization and NOAB construction procedures are provided in Tables 1 and 2, respectively.

For notational convenience, let $C(x)$ denote one of three constraint sets for the MIP: $C(x) = 1$ for continuous factors, 2 for discrete factors, and 3 for categorical factors, respectively. The inputs to our procedure are provided in Table 1, along with an initial balance feasibility test based on the discrete and categorical factors. This ensures that n is sufficiently large that the minimum analytically achievable imbalance does not exceed our criteria δ, and avoids situations like some in Figure 1 where no design will be able to satisfy the nearly balanced criterion. The requirement that $0 \leqslant \delta^* < 1$ in Table 1 ensures that every level of every factor appears at least once—an important feature for trade-off analyses.

Once the initial balance feasibility is confirmed, we consider orthogonality. The pairwise correlation between columns x and y of M is:

$$\rho_{xy} = \frac{\sum_{r=1}^{n} (m_{rx} - \bar{x})(m_{ry} - \bar{y})}{s_x s_y}. \tag{3}$$

If we fix all columns of M except column x, this means that the m_{ry}, $\bar{y}$, and s_y are all constants for $y \neq x$. Define

$$\rho_{xy}^* = \rho_{xy} s_x = \frac{\sum_{r=1}^{n} (m_{rx} - \bar{x})(m_{ry} - \bar{y})}{s_y}. \tag{4}$$

If x is balanced, the order in which its components are arranged does not affect its standard deviation s_x. This means that $\rho^*_{xy} \propto \rho_{xy}$, and that optimizing $z^* = arg\ \min \max_{y \neq x} |\rho^*_{xy}|$ is equivalent to optimizing $z = arg\ \min \max_{y \neq x} |\rho_{xy}|$. Although mathematical programming approaches cannot deal directly with this form of an objective function, we can define a quantity v and constrain it to satisfy $v \geqslant \max_{y \neq x} \rho^*_{xy}$ and $v \geqslant -\max_{y \neq x} \rho^*_{xy}$. With suitable constraints, one can then optimize v as a linear function of the entries in x. If the resulting $v = 0$, then x is orthogonal to all other columns in M. A MIP formulation is required because integer-valued variables are used in the design construction process. This approach was taken in Vieira *et al* (2011a).

For nearly balanced designs with discrete factors, Equation (4) does not decouple so easily. Let $\tilde{\mathcal{B}}(n, \ell)$ denote the set of nearly balanced vectors of length n with levels 1, …, ℓ. Then as long as the maximum imbalance δ is small, for any x_i, $x_j \in \tilde{\mathcal{B}}(n, \ell)$ we have $s_{x_i} \approx s_{x_j}$. Thus, $\rho^*_{xy} \tilde{\propto} \rho_{xy}$, where $\tilde{\propto}$ means approximately proportional. As in Hernandez *et al* (2012) and Vieira *et al* (2011a), with suitable constraints we can optimize v as a linear function of the entries in x. We no longer require the solutions to achieve $v = 0$, but merely that M retains its near-orthogonality property after the addition of column x. A similar approach is used for categorical factors, although it requires us to move to an indicator factor characterization and modify some constraints.

$\tilde{\mathcal{B}}(n, \ell)$ can be very large, which makes a formulation even more advantageous than random search. Our MIP starts with a feasible vector to work from at each step. These can easily be generated: include $\lfloor n/\ell \rfloor$ of each level, and then add the remaining values by sampling without replacement from a set that contains $(\lceil n/\ell \rceil - \lfloor n/\ell \rfloor)$ instances of each level.

Even if the balance feasibility check of Table 1 passes for a specified combination of δ^*, α^*, and n, this does not imply that the NOAB construction method will necessarily find a solution. Our experience shows that it is best to build the design beginning with categorical or discrete factors with low numbers of levels, and then add the continuous factors.

8. Large-study designs and performance

Our motivation for creating these nearly balanced, nearly orthogonal mixed designs arose from needs for large-scale simulation studies in a variety of application areas related to defense and national security. Rather than provide details about the factors, settings, results, and interpretation for any single study, we now provide brief descriptions of the design characteristics for some recent simulation experiments.

8.1. A flexible, customizable design for numerical factors

To allow analysts to quickly construct a customized design for a variety of settings without having to use the iterative MIP solution method, we constructed a 'general' big design with 512 design points. It can handle up to 100 continuous-valued factors and 10 blocks of 20 k-level discrete-valued factors, $k = 2,3,\ldots,11$, for a total of 300 quantitative factors, with $\delta = 0.1133$ and $\rho_{map} = 0.0356$. This simplifies the implementation of efficient, large-scale simulation experiments. An analyst with a problem involving up to 100 continuous factors and up to 20 k-level discrete factors, $k \in \{2,3,\ldots,11\}$, can quickly construct a design without having to implement our MIP procedure, simply by selecting the columns corresponding to his or her problem levels. For example, a design for 12 continuous factors, 3 two-level discrete factors and 7 nine-level discrete factors can be constructed from the first 12 continuous columns, the first 3 two-level columns, and the first 7 nine-level columns of our NOAB(300, 512) design, resulting in $\rho_{map} = 0.0252$ and $\delta = 0.1035$. Two-level qualitative factors can also be accommodated with this design.

There are few competitors for the NOAB(300, 512) design. A two-level fractional factorial can be constructed, but this has extremely poor balance ($\delta = 255.0$) because it samples only at the low and high levels of the factor range; central composite and D-optimal designs have similar problems. As Figure 2 demonstrates, constructing space-filling designs and then rounding them does not work as well as one might expect. We constructed six different space-filling designs: a maximin LH design, a sphere packing design (Johnson *et al*, 1990), a maximum entropy design (Shewry and Wynn, 1987), a minimum potential design, as well as a Sobol and scrambled Sobol sequence (Sobol, 1967). Rounding these to obtain designs comparable to our NOAB (300 512), we find that they have a wide range of performance in terms of balance (δ ranging from 0.033 for maximin LH to 248.5 for sphere-packing), but none came close to the NOAB in terms of correlation (ρ_{map} ranging from 0.168 for sphere-packing to 0.965 for the Sobol sequence design). We did not construct a Uniform design (Fang and Wang, 1994) because the estimated time required was over 25 234 CPU hours.

Our NOAB(300, 512) has already been used for several studies. As one example, it was used extensively over a 1-month period as part of the model verification stage to determine whether new features implemented in a Cultural Geography (CG) simulation were working properly. The CG model was developed by the US Army to explore how potential responses of different demographic segments of population in the Khandahar province of Afghanistan to various interventions and activities conducted by the International Security Assistance Force or insurgent groups. In all, over 30 000 runs were conducted in a 4-week period. High-performance computing assets were used extensively, but even so it could take up to 48 h to run 5 replications of our NOAB design. We conducted a series of experiments involving between 15 and 45 factors of mixed types and levels. The ability to adapt the design quickly in order to add additional factors proved to be invaluable. The space-filling nature of this design was also extremely important, because it tested the software at a wide variety of input setting combinations. This uncovered model bugs or anomalous behaviour that might have not been evident with more limited testing, but needed to be addressed before moving forward. Although our primary purpose in this initial stage was to assist in model

verification during this rapid model development cycle, the ultimate goal was to conduct large-scale experiments involving the CG model to examine trade-offs among various types of interventions. For example, understanding the costs and benefits of various combinations of infrastructure development projects (such as schools, medical facilities, irrigation etc) may improve the Army's ability to make effective use of limited budgets.

8.2. Customized designs for numerical and qualitative factors

In addition to the general-purpose NOAB(300, 512) design, we have constructed customized designs for a number of different applications. As a rule of thumb, we have found that when constructing designs involving K quantitative factors, designs of size n with $3K \leqslant n \leqslant 10K$ provide a good mix of efficiency, statistical power, and analysis flexibility. Similarly, for designs where L of the K factors are qualitative, we generally seek designs of size n with $3\left(K-L+\sum_{x\in L}(\ell_x-1)\right) \leqslant n \leqslant 10\left(K-L+\sum_{x\in L}(\ell_x-1)\right)$.

One recent example is a design used in a case study under development by one of NATO Modeling and Simulation Group's technical activities. This involves the protection of a small combat outpost against an insurgent attack. There are 21 factors related to equipment, soldier capabilities, and fire support that could be requested in the event of an attack. Nine of these are discrete with various low numbers of levels, and seven factors are categorical with three or more levels. The NOAB design had 168 design points, $\rho_{map} = 0.0138$, $\delta = 0.142$, and was used to explore trade-offs related to the formulation and effectiveness of various force protection strategies. This case study highlights the benefits of using large-scale designed experiments to gain operational insights from simulation models, and is intended to serve as a proof-of-concept demonstration about these benefits to the broader NATO community.

We have also constructed customized designs for a variety of other trade-off analyses, including applications to air reconnaissance missions, unmanned aerial vehicle capabilities, logistics life cycle management, and other topics in defense and national security. We refer the reader to Sanchez *et al* (2012) for a more detailed illustration of the types of analyses that can be conducted.

9. Small-study performance comparisons

NOAB designs were initially intended for usage in large-scale simulation experiments. They are well suited for trade-off analyses involving large numbers of factors and levels of interest. We have shown that alternative designs for trade-off analyses are not readily available, and described a few of many studies that have already benefitted from the use of NOAB designs. However, other designs are available if we are studying simpler systems (or restrict our attention to a limited number of factors *a priori*). We now show that even in these more limited scenarios, this approach can produce designs that outperform existing nearly-orthogonal alternatives for discrete factors.

In addition to ρ_{map}, some criteria often used to evaluate designs for small-scale experiments are the number of non-orthogonal pairs, along with the $\boldsymbol{D}$-optimality and $\boldsymbol{D}$-efficiency, typically computed using a main-effects response model assumption. In our experience, the complexity of responses for trade-off studies means this assumption is inappropriate. Nonetheless, calculating the $\boldsymbol{D}$-optimality and $\boldsymbol{D}$-efficiency for our designs does provide a bound on how well (or poorly) our designs would perform relative to the designs that yield the most precise effect estimates in these idealized situations.

Wang and Wu (1992) compared their NOAs with those of Taguchi (1959) and Tukey (1959). Later, Nguyen and Liu (2008) did the same with the proposals of Wang and Wu (1992), Ma *et al* (2000), and Xu (2002), producing arrays that outperform the previous ones.

We compare our NOAB designs with those found by Nguyen and Liu (2008) in Table 3. In their work, Nguyen and Liu (2008) created designs for categorical data, so they calculated $\boldsymbol{D}$-optimality as $|R|^{1/k}$, where R is the correlation matrix of the K columns of the matrix of orthogonal polynomial contrasts and $|R|$ is its determinant. As we aim to construct designs for numerical data, we also calculate the $\boldsymbol{D}$-optimality using the original design matrices. We compute the number of non-orthogonal pairs and ρ_{map} for both the categorical and numerical approaches. We also present the $\boldsymbol{D}$-efficiency$\equiv(1/n)|\mathbf{M}'\mathbf{M}|^{1/k}$ for the numerical approaches.

The best results in each of the criteria for numerical designs are highlighted in bold in Table 3. In 16 of the 21 cases, our proposal produces as good or better results in all five criteria (number of non-orthogonal pairs, δ, ρ_{map}, $\boldsymbol{D}$-optimality, and $\boldsymbol{D}$-Efficiency) than the benchmark. Of the five exceptions, the two smallest designs are outperformed only in terms of imbalance. The other exceptions are the $L'_{18}(9{\cdot}2^8)$, $L'_{24}(3{\cdot}2^{21})$, and $L'_{24}(6{\cdot}2^{18})$, where our approach still produces the best results in at least *three* of the five criteria. For $L'_{18}(9{\cdot}2^8)$, our design is slightly imbalanced and underperforms the best design by very small amount (0.003) in $\boldsymbol{D}$-optimality. In the last two exceptions, we suggest that decreasing ρ_{map} from 0.408 to 0.102 (or 0.333 to 0.048) and increasing the $\boldsymbol{D}$-Efficiency by 0.024 or 0.027 is well worth the increase in the number of non-orthogonal pairs.

Our NOAB designs in Table 3 do not satisfy the definition of an OA, so we advocate their use only for discrete quantitative factors, not for qualitative factors. However, for a fixed number of design points, analysts can study a greater number of quantitative factors with our designs than by using existing OAs. Conversely, for a fixed number of factors, our designs often require far fewer design points.

10. Balance, orthogonality, and space-filling comparisons for small studies

Lu *et al* (2013) discuss recent work on methods for directly constructing designs that perform well for multiple criteria, and

Table 3 Comparison of small nearly orthogonal designs using five criteria (best instances in bold)

	Nguyen and Liu (2008) Proposal								*NOAB Designs*				
Design	*Categorical**			*Numerical*					*Numerical*				
	$\#_{no}$	ρ_{map}	D_{opt}	$\#^N_{no}$	δ	ρ^N_{map}	D^N_{opt}	D^N_{eff}	$\#^N_{no}$	δ	ρ^N_{map}	D^N_{opt}	D^N_{eff}
$L'_6(3{\cdot}2^3)$	3	0.333	0.901	3	**0**	**0.333**	0.877	0.830	**1**	0.333	**0.333**	**0.971**	**0.880**
$L'_{10}(5{\cdot}2^5)$	10	0.200	0.967	10	**0**	0.200	0.951	**0.867**	**6**	0.2	**0.166**	**0.976**	**0.867**
$L'_{12}(4{\cdot}3^4)$	6	0.250	0.946	6	**0**	0.250	0.973	0.676	**0**	**0**	**0.000**	**1.000**	**0.692**
$L'_{12}(2^3{\cdot}3^4)$	6	0.250	0.946	6	**0**	0.250	0.980	0.802	**0**	**0**	**0.000**	**1.000**	**0.816**
$L'_{12}(6{\cdot}2^5)$	4	0.333	0.959	4	**0**	0.333	0.933	0.847	**1**	**0**	**0.098**	**0.998**	**0.896**
$L'_{12}(6{\cdot}2^6)$	6	0.333	0.947	6	**0**	0.333	0.918	0.843	**2**	**0**	**0.097**	**0.997**	**0.907**
$L'_{12}(3{\cdot}2^9)$	8	0.333	0.933	8	**0**	0.333	0.927	0.900	**4**	**0**	**0.204**	**0.982**	**0.948**
$L'_{12}(2{\cdot}3^5)$	10	0.250	0.877	10	**0**	0.250	0.930	0.704	**0**	**0**	**0.000**	**1.000**	**0.749**
$L'_{12}(3^2{\cdot}2^7)$	8	0.333	0.888	**6**	**0**	0.333	0.871	0.815	7	**0**	**0.204**	**0.965**	**0.893**
$L'_{12}(3^3{\cdot}2^5)$	9	0.333	0.925	6	**0**	0.333	0.926	0.816	**3**	**0**	**0.125**	**0.994**	**0.869**
$L'_{15}(5{\cdot}3^5)$	10	0.200	0.882	10	**0**	0.200	0.959	0.654	**0**	**0**	**0.000**	**1.000**	**0.678**
$L'_{18}(2{\cdot}3^8)$	3	0.289	0.967	2	**0**	0.250	0.985	0.713	**0**	**0**	**0.000**	**1.000**	**0.723**
$L'_{18}(2^3{\cdot}3^7)$	3	0.333	0.970	**3**	**0**	0.333	0.949	0.737	**3**	**0**	**0.111**	**0.996**	**0.770**
$L'_{18}(9{\cdot}2^8)$	28	0.111	0.985	28	**0**	0.111	0.973	**0.894**	**21**	0.111	**0.100**	**0.982**	**0.894**
$L'_{20}(5{\cdot}2^{15})$	18	0.200	0.956	18	**0**	0.200	0.947	0.910	**7**	**0**	**0.070**	**0.997**	**0.958**
$L'_{24}(8{\cdot}3^8)$	28	0.125	0.897	28	**0**	0.125	0.973	0.648	**0**	**0**	**0.000**	**1.000**	**0.664**
$L'_{24}(3{\cdot}2^{21})$	8	0.333	0.968	**6**	**0**	0.408	0.966	0.951	15	0.125	**0.102**	**0.992**	**0.975**
$L'_{24}(6{\cdot}2^{15})$	1	0.333	0.994	1	**0**	0.333	0.993	0.950	**0**	**0**	**0.000**	**1.000**	**0.956**
$L'_{24}(6{\cdot}2^{18})$	6	0.333	0.974	**6**	**0**	0.333	0.969	0.934	11	**0**	**0.048**	**0.999**	**0.961**
$L'_{24}(2{\cdot}3^{11})$	55	0.177	0.900	50	**0**	0.204	0.950	0.677	**6**	**0**	**0.062**	**0.997**	**0.707**
$L'_{24}(3{\cdot}4^7)$	21	0.236	0.872	19	**0**	0.200	0.972	0.590	**0**	**0**	**0.000**	**1.000**	**0.605**

**Source*: Nguyen and Liu (2008, p. 5274). $\#_{no}$: categorical number of non-orthogonal pairs; ρ_{map}: categorical maximum absolute pairwise correlation; D_{opt}: categorical *D*-optimality; $\#^N_{no}$: numerical number of non-orthogonal pairs; δ: maximum imbalance; ρ^N_{map}: numerical maximum absolute pairwise correlation; D^N_{opt}: numerical *D*-optimality; D^N_{eff}: numerical *D*-efficiency.

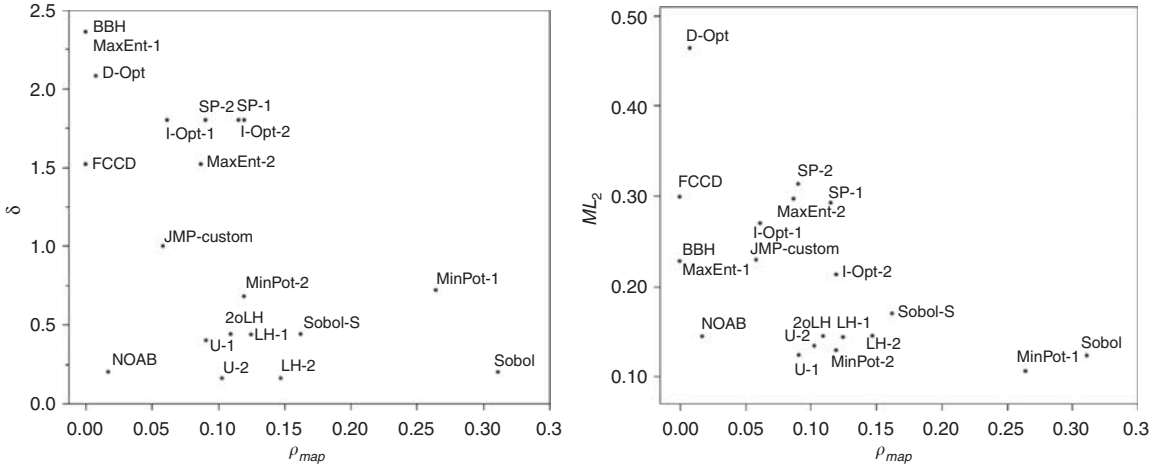

Figure 3 δ and ML_2 *versus* ρ_{map} for 20 designs for S^{20}.

propose methods for identifying designs that lie on the Pareto frontier. They also provide examples for some small-scale studies for applications where the total number of observations is extremely limited, and examine the robustness of the design choices to the weights assigned to different criteria. While this approach of generating designs does not lend itself to large-scale simulation studies, Pareto frontiers are useful ways of comparing designs.

We follow the study by MacCalman *et al* (2013). In that study, they propose a second-order NOLH and compare its space-filling performance and orthogonality with that of the following four traditional designs: the Faced Central Composite Design (Myers *et al*, 2009), BBH design (Box and Behnken, 1960), D-Optimal, and I-Optimal designs, as well as several space-filling designs. They used the SAS Institute's JMP software (see http://www.jmp.com) to create each of the

alternative designs for their comparison by instantiating the algorithm 500 times for the Sphere Pack, Uniform, and LH designs, and 30 times for the D-Optimal, I-Optimal, and Max Entropy designs and selecting the one with the lowest maximum pairwise correlation among any two columns in a second-order analysis matrix. They used as space-filling measure the modified L2 discrepancy (ML_2) (given by 5), which assesses how well a design covers the entire design region; the smaller the value, the better a design's space-filling property (Hickernell, 1998).

$$ML_2 = \left(\frac{4}{3}\right)^k - \frac{2^{1-k}}{n}\sum_{d=1}^{n}\prod_{i=1}^{k}\left(3-x_{di}^2\right) + \frac{1}{n^2}\sum_{d=1}^{n}\sum_{j=1}^{n}\prod_{i=1}^{k}\left[2-\max\left(x_{di}, x_{ji}\right)\right] \quad (5)$$

MacCalman *et al* (2013) consider only continuous-valued factors, but we assume that our experiment has four discrete factors with 3, 4, 5, and 7 levels, respectively, and apply appropriate rounding. We also examine the designs (from JMP) chosen as best out of 500 starts for each of the 6 space-filling design approaches. This gives us 19 alternatives to our NOAB design, each having 25 design points and 4 numerical factors. These are compared with our NOAB design in Figure 3 in terms of ρ_{map}, δ, and ML_2. As the reader can see, the NOAB outperforms all the others in at least one criteria. It is on the Pareto frontier, and the nearest to the lower left corner.

11. Conclusions

In this paper, we provided a mixed-integer programming formulation that allows us to construct efficient, nearly orthogonal, nearly balanced designs for mixed factor problems. We call these NOAB designs. Our focus is on efficient NOABs that will be useful for practical applications involving trade-off analyses—that is, designs suitable for *large numbers* of factors where the factors may not all have the same number of levels.

For a specified number of design points n, our approach generates a design that is nearly orthogonal and also nearly balanced for any mix of factor types (categorical, numerical discrete, and numerical continuous) and/or number of factor levels. This can be used to create designs with low maximum absolute pairwise correlation, along with low maximum imbalance, for large-scale simulation problems involving any type of factors. The designs we construct require orders of magnitude fewer design points than many other approaches.

These new designs greatly expand the portfolio of designs available for analysts conducting large-scale simulation experiments. Consequently, there are much greater opportunities for gaining insights about the behaviour of complex simulation models—and the real-world situations they represent—in a timely manner. This may be particularly valuable for those who use simulation experiments as a quantitative basis for trade-off analyses. An advantage of these designs is that they are amenable to analysis using a variety of methods, including polynomial model-fitting, partition tree analysis, graphical methods, and more. The ability to handle large numbers of factors of different types is beneficial for robust analysis as well.

Our large NOAB is available online, providing off-the-shelf flexibility for analysts when designing a large-scale experiment. We hope that this and other designs will become more readily used, and so improve the insights that simulation studies can provide to decision makers.

Acknowledgements—This work was supported in part by grants from the Office of Naval Research and the U.S. Army Survivability and Lethality Analysis Directorate, and a grant of computer time from the DOD High Performance Computing Modernization Program at the Navy DSRC at the Stennis Space Center.

References

Atkinson AC and Donev A (1992). *Optimum Experimental Designs*. Oxford University Press: Oxford, UK.

Bathke A (2004). The ANOVA F test can still be used in some balanced designs with unequal variances and nonnormal data. *Journal of Statistical Planning and Inference* **126**(2): 413–422.

Box GEP and Behnken DW (1960). Some new three level designs for the study of quantitative variables. *Technometrics* **2**: 455–475.

Cioppa TM and Lucas TW (2007). Efficient nearly orthogonal and space-filling Latin hypercubes. *Technometrics* **49**(1): 45–55.

Dellino G, Kleijnen JPC and Meloni C (2012). Robust optimization in simulation: Taguchi and Krige combined. *INFORMS Journal on Computing* **24**(3): 471–484.

Fang KT and Wang Y (1994). *Number–Theoretic Methods in Statistics*. Chapman and Hall: London, UK.

Fisher RA (1925). *Statistical Methods for Research Workers*. Oliver and Boyd: Edinburgh.

Florian A (1992). An efficient sampling scheme: Updated Latin hypercube sampling. *Probabilistic Engineering Mechanics* **7**(2): 123–130.

Hedayat AS, Sloane NJ and Stufken J (1999). *Orthogonal Arrays: Theory and Applications*. Springer-Verlag: New York.

Hernandez AS, Lucas TW and Carlyle M (2012). Enabling nearly orthogonal Latin hypercube construction for any non-saturated run-variable combination. *ACM Transactions on Modeling and Computer Simulation* **22**(4): 20: 1–20: 17.

Hickernell FJ (1998). A generalized discrepancy and quadrature error bound. *Mathematics of Computation* **67**: 299–322.

Johnson ME, Moore L and Ylvisaker D (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference* **26**(2): 131–148.

Joseph VR and Hung Y (2008). Orthogonal-maximin Latin hypercube designs. *Statistica Sinica* **18**(1): 171–186.

Kleijnen JPC (2007). *Design and Analysis of Simulation Experiments*. Springer: New York.

Kleijnen JPC (2008). Simulation experiments in practice: Statistical design and regression analysis. *Journal of Simulation* **2**(1): 19–27.

Kleijnen JPC, Sanchez SM, Lucas TW and Cioppa TM (2005). A user's guide to the brave new world of designing simulation experiments. *INFORMS Journal on Computing* **17**(3): 263–289.

Kuhfeld WF (2010). Orthogonal Arrays, SAS. http://support.sas.com/techsup/technote/ts723.html, accessed 24 May 2012.

Law AM (2007). *Simulation Modeling and Analysis*. 4th edn. MacGraw-Hill: New York, USA.

Lu L, Anderson-Cook CM and Robinson TJ (2013). Optimization of designed experiments based on multiple criteria using a Pareto frontier. *Technometrics* **53**(4): 353–365.

Ma C, Fang KT and Liski E (2000). A new approach in constructing orthogonal and nearly orthogonal arrays. *Metrika* **50**(3): 255–268.

MacCalman AD, Vieira Jr H and Lucas TW (2013). *Second order nearly orthogonal modified Latin hypercubes for exploring models with multiple unknown response surface forms*. Working paper, under review.

Montevechi JAB, de Almeida Filho RG, Paiva AP, Costa RFS and Medeiros AL (2010). Sensitivity analysis in discrete-event simulation using fractional factorial designs. *Journal of Simulation* **4**(2): 128–142.

Montgomery DC (2005). *Design and Analysis of Experiments*. 6th edn John Wiley & Sons: New York.

Myers RH, Montgomery DC and Anderson-Cook CM (2009). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. 3rd edn. John Wiley & Sons: New York, USA.

Nguyen N-K and Liu M-Q (2008). An algorithmic approach to constructing mixed-level orthogonal and near-orthogonal arrays. *Computational Statistics and Data Analysis* **52**(12): 5269–5276.

Ryan TP (1997). *Modern Regression Methods*. John Wiley & Sons: New York.

Sacks J, Welch WJ, Mitchell TJ and Wynn HP (1989). Design and analysis of computer experiments (includes comments and rejoinder). *Statistical Science* **4**(4): 409–435.

Sanchez SM, Lucas TW, Sanchez PJ, Nannini CJ and Wan H (2012). Designs for large-scale simulation experiments, with applications to defense and homeland security. In: Hinkelmann K (ed). *Design and Analysis of Experiments*. Vol. 3, Wiley: New Jersey.

Sanchez SM and Wan H (2012). Work smarter, not harder: A tutorial on designing and conducting simulation experiments. In: Laroque C, Himmelspach J, Pasupathy R, Rose O, and Uhrmacher A M (eds), *Proceedings of the 2012 Winter Simulation Conference*, pp 1929–1943.

Santner TJ, Williams BJ and Notz WI (2003). *The Design and Analysis of Computer Experiments*. Springer-Verlag: New York.

Shewry MC and Wynn HP (1987). Maximum entropy sampling. *Journal of Applied Statistics* **14**(2): 165–170.

Sloane NJ (2007). Orthogonal Arrays. http://neilsloane.com/oadir/, accessed 15 September 2013.

Sobol IM (1967). Distribution of points in a cube and approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics* **7**(4): 86–112.

Taguchi G (1959). Linear graphs for orthogonal arrays and their applications to experimental designs, with the aid of various techniques. *Reports of Statistical Applications Research, Japanese Union of Scientists and Engineers* **6**: 1–43.

Tukey J (1959). Little pieces of mixed factorials. Unpublished manuscript.

Van Vorst N, Erazo M and Liu J (2012). PrimoGENI for hybrid network simulation and emulation experiments in GENI. *Journal of Simulation* **6**(3): 179–192.

Vieira Jr H (2011). *Selecting the system most likely to be the best in the presence of an infinite number of alternatives*. PhD Thesis, Aeronautics Institute of Technology. http://www.bd.bibl.ita.br/tesesdigitais/62037.pdf, accessed 24 May 2012.

Vieira Jr H, Sanchez S, Kienitz KH and Belderrain MCN (2011a). Generating and improving orthogonal designs by using mixed integer programming. *European Journal of Operational Research* **215**(3): 629–638.

Vieira Jr H, Sanchez SM, Kienitz KH and Belderrain MCN (2011b). Improved eficient, nearly orthogonal, nearly balanced mixed designs. In: Jain S, Creasey RR, Himmelspach J, White KP, and Fu M (eds), *Proceedings of the 2011 Winter Simulation Conference*, pp 3605–3616.

Wang JC and Wu CFJ (1992). Nearly orthogonal arrays with mixed levels and small runs. *Technometrics* **34**(4): 409–421.

Xu H (2002). An algorithm for constructing orthogonal and nearly-orthogonal arrays with mixed levels and small runs. *Technometrics* **44**(4): 356–368.

Appendix

MIP formulation and constraint sets $C(x)$

In this Appendix, we provide the MIP formulation and constraint sets. We assume the inputs have already been chosen so the initial balance feasibility test is passed, as in Table 1. The MIP is called with the inputs $\tilde{\mathrm{M}}_j$, δ^*, x, ℓ_x, and $C(x)$. Note that $\tilde{\mathrm{M}}_j = \mathrm{M}_j$ if no categorical factors have previously been added to the design. Otherwise, $\tilde{\mathrm{M}}_j$ will include appropriate indicator variables for these factors.

Some additional notation is needed. Let $I(i)$ denote the set of integers $\{1,2, \ldots, i\}$. Let x_r denote the value of the rth row of the new column x. ℓ_x is the number of levels of column x, and $\pi = \{\pi_1, \pi_2, \ldots, \pi_{\ell_x}\}$ is the set of the ℓ_x levels that the factor x can assume. The binary decision variables θ_{rl} have value 1 if $x_r = \pi_l$, and 0 otherwise.

The formulation for $C(x) = 1$ is adapted from Vieira *et al* (2011a) (see also Vieira, 2011). We do not allow imbalance for continuous factors, because that would mean one or more levels were never investigated. The formulations for $C(x) = 2$ and $C(x) = 3$ are adapted from Vieira *et al* (2011b), but we modify the imbalance constraints (v) and (vi) to accommodate a slightly different definition of imbalance.

FORMULATION for $C(x) = 1$: Adding a single continuous factor

Minimize v

Subject to

$$(i) \quad v \geqslant \frac{1}{s_c} \sum_{r=1}^{n} \left(x_r - \frac{1}{n} \sum_{k=1}^{n} x_k \right) (m_{rc} - \overline{c}) \qquad c \in I(j)$$

$$(ii) \quad v \geqslant -\frac{1}{s_c} \sum_{r=1}^{n} \left(x_r - \frac{1}{n} \sum_{k=1}^{n} x_k \right) (m_{rc} - \overline{c}) \qquad c \in I(j)$$

$$(iii) \quad \sum_{j=1}^{n} \theta_{rj} = 1 \qquad r \in I(n)$$

$$(iv) \quad x_r = \sum_{\ell=1}^{n} \pi_\ell \theta_{r\ell} \qquad r \in I(n)$$

$$(v) \quad \theta_{r\ell} \in \{0, 1\} \qquad r \in I(n); \ell \in I(n)$$

Constraint sets (i) and (ii), together with the minimization of v, ensure that $v = |\rho^*_{xy}|$. Constraint set (iii) ensures that only one level is assigned to each row of x. Constraint set (iv) converts the binary decision variables to their corresponding integer-valued levels.

FORMULATION for $C(x) = 2$: adding a single discrete factor

Minimize v

Subject to

$$(i) \quad v \geqslant \frac{1}{s_c}\sum_{r=1}^{n}\left(x_r - \frac{1}{n}\sum_{k=1}^{n}x_k\right)(m_{rc} - \overline{c}) \qquad c \in I(j)$$

$$(ii) \quad v \geqslant -\frac{1}{s_c}\sum_{r=1}^{n}\left(x_r - \frac{1}{n}\sum_{k=1}^{n}x_k\right)(m_{rc} - \overline{c}) \qquad c \in I(j)$$

$$(iii) \quad \sum_{\ell=1}^{\ell_x}\theta_{r\ell} = 1 \qquad r \in I(n)$$

$$(iv) \quad x_r = \sum_{\ell=1}^{\ell_x}\pi_\ell\theta_{r\ell} \qquad r \in I(n)$$

$$(v) \quad \sum_{r=1}^{n}\theta_{r\ell} \leqslant \left\lfloor (1+\delta)\frac{n}{\ell_x}\right\rfloor \qquad \ell \in I(\ell_x)$$

$$(vi) \quad \sum_{r=1}^{n}\theta_{r\ell} \geqslant \left\lceil (1-\delta)\frac{n}{\ell_x}\right\rceil \qquad \ell \in I(\ell_x)$$

$$(vii) \quad \theta_{r\ell} \in \{0,\ 1\} \qquad r \in I(n); \ell \in I(\ell_x)$$

Constraint sets (*i*) and (*ii*), together with the minimization of v, ensure that $v = |\rho^*_{xy}|$. Constraint set (*iii*) ensures that only one of the ℓ_x levels is assigned to each row of x. Constraint set (*iv*) translates the binary decision variables to their corresponding integer-valued levels. Constraint sets (*v*) and (*vi*) enforce the balance property.

FORMULATION for $C(x) = 3$: Adding a single categorical factor

Minimize v

Subject to

$$(i) \quad v \geqslant \frac{1}{s_c}\sum_{r=1}^{n}\left(x^i_r - \frac{1}{n}\sum_{k=1}^{n}x^i_k\right)(m_{rc} - \overline{c}) \qquad c \in I(j); i = I(\ell_x - 1)$$

$$(ii) \quad v \geqslant -\frac{1}{s_c}\sum_{r=1}^{n}\left(x^i_r - \frac{1}{n}\sum_{k=1}^{n}x^i_k\right)(m_{rc} - \overline{c}) \qquad c \in I(j); i = I(\ell_x - 1)$$

$$(iii) \quad \sum_{\ell=1}^{3}\theta^i_{r\ell} = 1 \qquad r \in I(n); i = I(\ell_x - 1)$$

$$(iv) \quad x^i_r = \sum_{\ell=1}^{3}(\ell - 2)\theta^i_{r\ell}, \qquad r \in I(n); i = I(\ell_x - 1)$$

$$(v) \quad \sum_{r=1}^{n}\theta^i_{r\ell} \leqslant \left\lfloor (1+\delta)\frac{n}{\ell_x}\right\rfloor \qquad \ell = 1, 3; i \in I(\ell_x - 1)$$

$$(vi) \quad \sum_{r=1}^{n}\theta^i_{r\ell} \geqslant \left\lceil (1-\delta)\frac{n}{\ell_x}\right\rceil \qquad \ell = 1, 3; i \in I(\ell_x - 1)$$

$$(vii) \quad \sum_{i=1}^{\ell_x - 1}\theta^i_{r3} \leqslant 1 \qquad r \in I(n)$$

$$(viii) \quad \sum_{i=1}^{\ell_x - 1}\theta^i_{r2} \leqslant \ell_x - 2 \qquad r \in I(n)$$

$$(ix) \quad \theta^i_{r1} - \theta^1_{r1} = 0 \qquad r \in I(n); i = 2,\ 3,\ \ldots\ , \ell_x - 1$$

$$(x) \quad \theta^i_{r\ell} \in \{0,\ 1\} \qquad r \in I(n);\ \ell \in I(3);\ i \in I(\ell_x - 1)$$

Constraint sets (*i*) and (*ii*), together with the minimization of v, ensure that $v = |\rho^*_{xy}|$. Constraint set (*iii*) ensures that only one of the ℓ_x levels is assigned to each row of x. Constraint set (*iv*) translates the binary decision variables to their corresponding integer-valued levels. The imbalance limits are guaranteed by constraint sets (*v*) and (*vi*); note that these are enforced only for non-zero values of the indicator variables. Constraints (*vii*)–(*ix*) are needed to construct the indicator variables properly. Specifically, (*vii*) ensures that no two indicator variables can have 1's assigned to the same row if they correspond to the same categorical factor, that is, that multiple assignments do not occur for a particular design point (row) of x. Constraints (*viii*) ensure that at least one level is assigned to every row of x. Finally, constraint (*ix*) ensures that all indicator variables associated with x have a '-1' value assigned to the same row.

Received 01 August 2012;
accepted 23 July 2013