Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

DESIGNING SIMULATION EXPERIMENTS WITH CONTROLLABLE AND UNCONTROLLABLE FACTORS

Christian Dehlendorff
Murat Kulahci
Klaus Kaae Andersen

Department of Informatics and Mathematical Modelling
Technical University of Denmark
Bygning 321, Richard Petersens Plads
Lyngby, DK-2800, DENMARK

ABSTRACT

In this study we propose a new method for designing computer experiments inspired by the split plot designs used in physical experimentation. The basic layout is that each set of controllable factor settings corresponds to a whole plot for which a number of subplots, each corresponding to one combination of settings of the uncontrollable factors, is employed. The caveat is a desire that the subplots within each whole plot cover the design space uniformly. A further desire is that in the combined design, where all experimental runs are considered at once, the uniformity of the design space coverage should be guaranteed. Our proposed method allows for a large number of uncontrollable and controllable settings to be run in a limited number of runs while uniformly covering the design space for the uncontrollable factors.

1 INTRODUCTION

With the current advances in computing technology, computer and simulation experiments are increasingly being used to study complex systems for which physical experimentation is usually not feasible. Our case study involves a discrete event simulation model of an orthopedic surgical unit. The discrete event simulation (DES) model describes the individual patient's progress through the system and has been developed in collaboration with medical staff at Gentofte University Hospital in Copenhagen. The unit undertakes both acute and elective surgery and performs more than 4,600 operative procedures a year. While the patients come from various wards throughout the hospital, the main sources of incoming patients are the four orthopedic wards or the emergency care unit.

The simulation model is implemented in Extend version 6 (Krahl 2002) on a Windows XP platform and controlled from a Microsoft Excel spreadsheet with a Visual Basic for application script. The model consists of 3 main modules: The wards and arrival, the operating facilities, and the recovery and discharge. Interaction with the surrounding hospital is for example modeled with simplified processes using the same resources as the processes in the surgical unit (occupying the resources) and with the patients entering and exiting the model. Operating rooms, recovery beds, wards and staff are included in the model. The average run time for simulating 6 months (with one week of warm-up) operations is around 7 minutes. Typical outcomes are waiting times, patient throughput and the amount of overtime.

The simulation model has two sources of noise coming from variations in the uncontrollable factors (a.k.a. environmental factors in physical experimentation) and from changes in the seed controlling the random number generation process embedded in the simulation model. The controllable factors are for example the number of operating rooms and the number of surgeons, whereas the uncontrollable factors may include for example the arrival rate of acute patients and the time required to clean the operating rooms.

In this type of application, several issues need to be considered. First, the controllable factors tend to be numerous and often discrete. Moreover a single experiment usually takes several minutes to run. Therefore a simple exhaustive method, where all possible combinations of the factor settings are considered, is often computationally infeasible due to the exponentially increasing number of factor combinations. Furthermore, the settings of the uncontrollable factors, e.g. the acute patient arrival rate or the duration of surgical procedures, are also of interest and must be determined as they may influence the outcome of the simulations and hence the robustness of the simulation analysis.

The paper is organized in the following manner: Section 2 introduces design of computer experiments and de-

978-1-4244-2708-6/08/$25.00 ©2008 IEEE

fines the performance measure for the designs. Section 3 describes the proposed design method and contrasts it with other methods. In section 4 opportunities for future research are presented. Finally the main conclusions are summarized in section 5.

2 DESIGN OF COMPUTER EXPERIMENTS

2.1 Literature Review

A general discussion on the issues regarding the design and analysis of computer experiments can be found in Sacks et al. (1989), Santner, Williams, and Notz (2003) and Fang, Li, and Sudjianto (2006). The outputs from the computer experiments are often considered to come from a deterministic computer code. In such experiments, the classical design of experiment methods such as replication is deemed to be redundant as replication of an experiment, for example, yields exactly the same result (see Santner, Williams, and Notz (2003) and Fang, Li, and Sudjianto (2006)).

Experiments based on a simulation model often involve some stochastic component; making the output also stochastic. Kleijnen (2008) discusses the design and analysis of simulation experiments which typically have some sort of noise in the output. Therefore these experiments differ from the deterministic computer experiments. Furthermore, a typical simulation application will have both controllable and uncontrollable (environmental) factors, which should be handled differently. In these applications the aim is to manipulate the controllable factors so that the system is insensitive (robust) to changes in the uncontrollable factors. As described by Kleijnen (2008) and Sanchez (2000) the solution's robustness needs to be considered in order to obtain applicable solutions in systems with uncontrollable factors. That is, a good solution needs to perform well over the entire range of uncontrollable factors.

The original concept of robustness in physical systems is often attributed to Taguchi (1987). Taguchi's methods involve an inner array for the controllable factors and an outer array for the uncontrollable factors. In simulation studies, Kleijnen (2008) suggests using a crossed design, e.g. combining a central composite design (CCD) for the controllable factors and a Latin Hypercube Design (LHD) for the uncontrollable factors. In a crossed design the same set of subplots is used for each whole plot. However, as we will show in this study, this may not be the most efficient way of running such experiments.

2.2 Simulation Model

Our basis is a discrete event simulation model generating output, $y = f(\mathbf{x_c}, \mathbf{x_e})$, for the settings for the s_c controllable factors, $\mathbf{x_c}$, and the settings for the s_e uncontrollable factors, $\mathbf{x_e}$. The objective is not only to select the settings, $\mathbf{x_c^*}$, such that the solution is robust to changes in the uncontrollable factor settings as described in p. 130-134 in Kleijnen (2008), but also to understand the variation coming from the changes in the uncontrollable factor settings.

Since little prior knowledge of both controllable and uncontrollable factors is available, we require that a good design is simultaneously uniform over the design space of the controllable and uncontrollable factors. In the following, we will assume that the uniform coverage of the design space of the controllable factors is already achieved and that we are only concerned with the uncontrollable factors.

Robustness studies in physical experimentation often involve split-plot designs (Montgomery 2005). We will therefore use similar terminology when robustness studies are performed using computer experiments. In classic split-plot designs, a set of experiments called whole-plots is designed so that for each whole-plot another set of experiments called subplots are run. In robustness studies, the settings of the controllable factors often constitute the whole-plots, whereas the settings of the uncontrollable factors constitute the subplots. In Table 1, a whole-plot corresponds to a row in which randomly selected combinations of settings for the uncontrollable factors are run. It should be noted that the randomization issue is irrelevant for computer experiments.

In the proposed method, each whole-plot corresponds to one combination of settings of the controllable factors (a row in Table 1), i.e. a total of n_c whole-plots are needed ($n_c = 5$ in Table 1). Each subplot (a column entry in any row in Table 1) corresponds to a combination of settings for the uncontrollable factors with a total of k subplots for each whole-plot. Thus the overall design consists of $N = n_c k$ runs. In a crossed design as proposed by Kleijnen (2008) these k subplots would be the same from one whole-plot to the next. Therefore there will only be a total of k combinations of settings for the uncontrollable factors. In our proposed methodology, different k combinations of settings for the uncontrollable factors will be used for each whole-plot. This is expected to give better overall coverage of the uncontrollable factor space compared to the crossed design. The challenge with the proposed method is to make the uncontrollable factor settings comparable from one whole-plot to the next.

2.3 Measure of Uniformity

In order to evaluate the designs presented in the following sections a measure of uniformity is needed. Fang, Li, and Sudjianto (2006) summarize a set of performance measures frequently used for measuring the uniformity of a design: the star discrepancy, centered discrepancy and the wrap-around discrepancy. The centered and the wrap-around discrepancy were proposed by Hickernell (1998b) and Hickernell (1998a), respectively.

Table 1: Uncontrollable factor design for five controllable settings and five environmental settings within each controllable setting

Controllable setting	Environmental setting				
	1	2	3	4	5
1	x_{e1}	x_{e2}	x_{e3}	x_{e4}	x_{e5}
2	x_{e6}	x_{e7}	x_{e8}	x_{e9}	x_{e10}
3	x_{e11}	x_{e12}	x_{e13}	x_{e14}	x_{e15}
4	x_{e16}	x_{e17}	x_{e18}	x_{e19}	x_{e20}
5	x_{e21}	x_{e22}	x_{e23}	x_{e24}	x_{e25}

Both have desirable properties. They are easy to compute, invariant to permutations of factors or runs and rotation of coordinates, and reliable measurements for the uniformity of projections. However the wrap-around discrepancy is said to be unanchored (i.e. it only involves the design points), while the centered discrepancy is not, since it involves the corners of the unit cube.

In this study only the wrap-around discrepancy is considered as the measure of uniformity with a low value corresponding to a high degree of uniformity. The measure is chosen since the literature generally suggests it as a good measure of uniformity (see for example Fang and Ma (2001); Fang, Lin, and Liu (2003); Fang, Li, and Sudjianto (2006)). The idea behind this measure is that for any two points from a uniform design, x_1 and x_2, spanning a hyper cube (potentially wrapping around the bounds of the unit cube); the hypercube should contain a fraction of the total number of points equal to the fraction of total volume covered by the cube. An analytic expression for the wrap-around discrepancy (WD(D)) is given by Fang and Ma (2001) as

$$(WD(D))^2 = -\left(\tfrac{4}{3}\right)^s + \tfrac{1}{n}\left(\tfrac{3}{2}\right)^s + \tfrac{2}{n^2}\sum_{k=1}^{n-1}\sum_{j=k+1}^{n}\prod_{i=1}^{s} d_i(j,k) \quad (1)$$

with $d_i(j,k) = \frac{3}{2} - |x_{ki} - x_{ji}|(1 - |x_{ki} - x_{ji}|)$, n being the number of points, s the number of factors (the dimension), and x_{ki} the i'th coordinate of the k'th point.

There are various ways of constructing uniform designs. In this study the good lattice point method based on the power generator is used with the modification described in Fang, Li, and Sudjianto (2006). The design construction is based on a lattice $\{1,\ldots,n\}$ and a generator $h(k) = (1,k,k^2,\ldots,k^{s-1})(\text{mod } n)$, with k fulfilling that $k,k^2,\ldots,k^{s-1}(\text{mod } n)$ are distinct. $h(k)$ is chosen such that the resulting design consisting of the elements $u_{ij} = ih(k)_j(\text{mod } n)$ scaled down to $[0,1]^s$ has the lowest WD-value.

3 DESIGN ALGORITHM

A method for generating good designs for simulation models with both controllable and uncontrollable factors is presented in the following section. Here we assume that all factors have been scaled to $[0,1]$ and that the wrap-around discrepancy is the measure of uniformity. It is furthermore assumed that a design for the controllable factors is available. That is, we are primarily concerned with designing experiments for the uncontrollable factors. Two and three dimensional examples are used since they can be illustrated graphically. However, the method is general and results for 4 and 10 factors are also presented.

3.1 Bottom-up Approach

In section 2.2 the limitations of crossing a design for the controllable factors with a design for the uncontrollable factors were described. A better method in terms of covering the uncontrollable factor space compared to the crossed design is to generate different designs for the whole-plots, each with k different combinations of uncontrollable factor settings. This implies that n_c designs of size k should be constructed. For this method to succeed in the combined design, not only sets of k subplots for different whole-plots should be comparable, but also $n_c k$ subplots need to cover the design space for the uncontrollable factors uniformly. This can be achieved by dividing the design hyperspace for the uncontrollable factors into k sub-regions and sample n_c settings in each. As shown in Figure 1, this can be achieved fairly easily in two dimensions. However, in higher dimensions an efficient way of generating the sub-regions is required since the curse of dimensionality dictates that exponentially increasing numbers of runs have to be used in higher dimensions to obtain the same density of runs as in the lower dimensions.

If regular partitioning of the hypercube is possible, a design can be generated by randomly taking a run from each sub-region for each whole-plot. Figure 1 illustrates the approach in two dimensions with 16 subplots in each of the 10 whole plots. The design in Figure 1 has poor overall uniformity, which can also be seen from WD-values being 12 to 51 times higher compared to a uniform design of the same size.

A general method for generating the sub-regions is to generate a uniform design of size k and use these points as center points of k hypercubes or spheres that will constitute the sub-regions. The subplots are then generated within these sub-regions by either uniform designs or maxi-min distance designs for which the minimum distance of two runs in a sub-region is maximized. Figure 2 illustrates the performance of these methods for five controllable and 40 environmental settings for two environmental factors. The performance parameter in the figure is the WD-value for

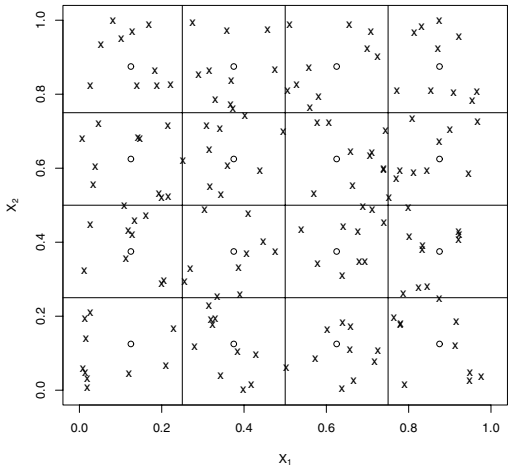

Figure 1: A total design of size 160 settings in 16 regions with 10 settings in each. Circles correspond to centers and crosses to sample settings.

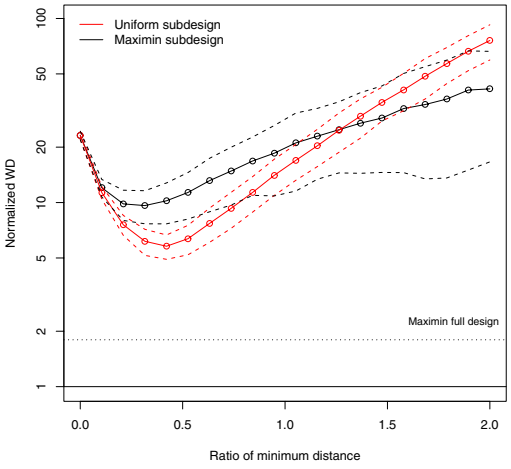

Figure 2: Average WD-value normalized using the WD-value obtained for a uniform design with 200 runs. Black curve with marks is for the maximum design and the red for the uniform design with dashed curves corresponding to approximate 95 % confidence intervals, the bottom black solid curve indicates a ratio of 1, i.e. no difference. The black dotted curve corresponds to a maxi-min distance. The overall design consists of 200 settings with the number of environmental settings being 40.

the combined environmental factor design, normalized by the WD-value of a uniform design of size 200. It can be seen that, compared to a uniform design generated directly for the same number of runs, both bottom-up methods are significantly worse. A maxi-min design generated directly is also seen to be better than the bottom-up generated designs. Figure 2 illustrates that using a bottom-up approach does not ensure an overall uniform design for the uncontrollable factors.

3.2 Top-down Approach

The second method we propose has more of a "top-down" structure. First, we generate a uniform design of size N which is equal to kn_c. This assures that the combined design is indeed uniform. But this does not solve the problem of assigning k settings to each of the n_c whole-plots such that in each whole-plot the subplots are uniformly spaced.

One approach to generate the designs is first to construct k sub-regions around k centers, where each region consists of n_c points. A method to obtain such a structure is to generate another uniform design of size k and use these points as starting center points, c, in an optimization algorithm that finds the optimal center points by minimizing

$$\sum_j \min_i ||x_j - c_i|| + k\sum_i (n_i - n_c)^2 \qquad (2)$$

In the above expression, n_i is the number of points having center i as the closest center. That is, the objective is to choose the centers, c^* such that they minimize the sum of the smallest differences between points and the centers, and the deviations from the required size of the region. This should ensure reasonably good separation of the points.

Based on the optimal centers, c^*, the N points need to be assigned to a center such that all points are assigned and all centers have exactly n_c points. This can be done in various ways, for example by assigning the point with the smallest distance to its nearest center, or by assigning the point with the largest second-shortest distance to its nearest center, or by simply considering the points' membership to each center based on euclidean distances.

A result of assigning 400 points to 10 groups of 40 points each is shown on the left of Figure 3, where it can be seen that the resulting groups are not well defined. Applying an exchange-algorithm on the assignment significantly improves the assignment as seen on the right of Figure 3. The total distances of the points to their center are reduced by 5 % by swapping less than 20 points and the points are grouped in well-defined clusters. An example in three dimensions is shown in Figure 4. The grouping in Figure 4 is generated by applying the exchange algorithm to a completely random assignment leading to a 49 % improvement

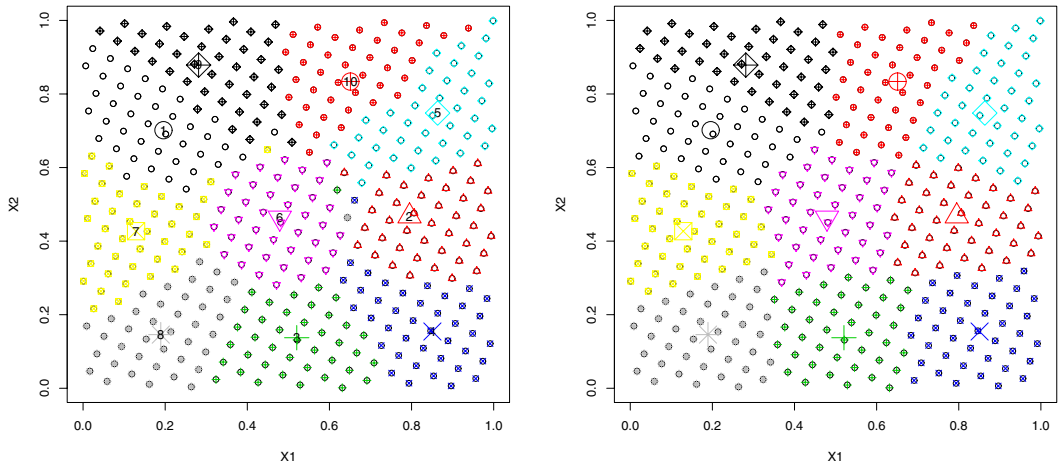

Figure 3: Left: The optimal assignment corresponding to a membership assignment. Right: The assignment after swapping in the optimal design.

in the distance of the points to the centers by more than 200 swaps.

3.2.1 Generating Whole Plots

After grouping the subplots in k groups, we generate the whole-plots. Each whole-plot is assigned to one setting from each of the k groups so that all settings are assigned. One method is to assign the settings such that the maximum WD-value of the sub-designs is minimized, which can be obtained by repeatedly assigning the settings randomly to the whole-plots until a certain degree of uniformity is obtained.

Another method is to move the small uniform design of size k so that the point closest to the origin in the small design is placed at the points in the group closest to the origin and then assign points based on the smallest distance. The advantage of this approach compared to random assignment is that the whole-plot approximately mimics the uniform design structure.

For the designs considered in Figure 3 and 4 the performance of each whole-plot is compared to a uniform design generated directly in Table 2. The table shows that the overall uniformity of the combined design cannot be fulfilled without getting sub-designs that are not completely uniform. The designs with lowest maximum relative WD-value all have WD-values below 3.7 times and the highest minimum WD-values are less than twice the reference designs.

It can be seen from Table 2 that the results are consistent for up to 10 factors. The mean and the smallest maximum WD-value are all decreasing, whereas the remaining values are inconclusive with respect to the number of factors. It can also be seen from Table 2 that a design, which ensures

Table 2: Summary for relative WD-values for 2, 3 and 4 dimensional examples with 40 controllable factors, each with 10 environmental settings (400) or 20 controllable factors, each with 10 environmental settings (200). The performance is summarized by minimum (Min), mean (Mean) and maximum (Max) relative WD-value and by the highest minimum (Max min) and lowest maximum (Min max). The values are relative to the WD-value for a uniform design of the same size as the whole-plots

Factors	Min	Max min	Mean	Min max	Max
2 (400)	1.15	1.99	2.78	3.67	8.39
3 (400)	1.19	1.93	2.70	3.47	7.21
4 (400)	1.25	1.94	2.56	3.20	7.28
10 (400)	1.32	1.60	1.76	2.00	2.38
2 (200)	1.14	2.17	2.69	2.94	7.20
3 (200)	1.17	2.21	2.68	2.94	6.98
4 (200)	1.22	2.22	2.50	2.54	5.65
10 (200)	1.29	1.63	1.73	1.78	2.45

relative WD-values for all whole-plots between 2 (Max min) and 3.7 (Min max) can be achieved for up to 10 factors. The results seem to be independent of the number of settings but with 10 factors generally giving significantly lower values. This may be caused by the sparsity of the settings in the 10 dimensional design space.

4 DISCUSSION

This study was originated from application of discrete event simulation and computer experimentation at a hospital unit.

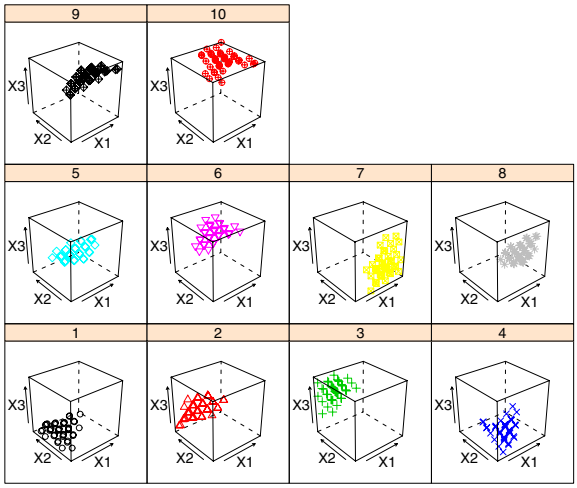

Figure 4: 400 settings assigned to 10 groups in 3 dimensions. Each panel corresponds to one group.

In health- care applications in general, it is desireable that the final solutions are robust to changes in the uncontrollable factors. In the proposed design a large set of combinations of the uncontrollable factor settings is achieved using only a limited number of runs in each whole-plot. This is due to the fact that in each whole-plot a different set of subplots is used. When considered together, however, the subplots in the combined design show a uniform coverage of the design space.

Based on the proposed design, a meta-model of the following form

$$y(x_e,x_c) = f_1(x_c) + f_2(x_e) + f_{12}(x_c,x_e) + e \qquad (3)$$

could be considered with $f_1(x_c)$ being a function describing the fixed effects related to the controllable setting, $f_2(x_e)$ and $f_{12}(x_c,x_e)$ being random effects describing the variations on the mean effect and the effect of the uncontrollable factor variations on the fixed effects.

By ensuring the overall uniformity of the uncontrollable factor settings, the functions $f_2(x_e)$ and $f_{12}(x_c,x_e)$ can be estimated over the whole region. The functions $f_2(x_e)$ and $f_{12}(x_c,x_e)$ describe the impacts of the variations in the uncontrollable factors. These can be used for quality improvement purposes if the variation in some of the uncontrollable factors is somehow possible to reduce. Moreover, $f_{12}(x_c,x_e)$ is of interest in robustness studies since the interaction between controllable and uncontrollable factors is the key to reducing the impact from changes in the uncontrollable factors.

5 CONCLUSION

In this study, a methodology to design uniformly distributed experiments for simulation experimentation in the presence of both controllable and uncontrollable factors is introduced. The method ensures that the subplots in the combined design for the uncontrollable factors are uniform while keeping an acceptable level of uniformity of the subplots within each whole-plot. Complete uniformity compared to uniform design of the size equal to the total number of subplots could not, however, be achieved.

The proposed methodology is primarily based on Euclidian distances. Therefore the method can be used in designs with many uncontrollable/environmental factors. Our results show that a uniformity measure of the individual whole-plots can be minimized to within two to four times the value of an overall uniform design. Furthermore, it was shown that the method was applicable to designs with 2 to 10 uncontrollable factors. Since the methodology is based on distances, increasing the number of factors may be possible, although sparsity of the experiments in the design space may become an issue.

The proposed design contains as many uncontrollable factor settings as the number of runs (N), which in contrast to a crossed design of the same size has $k = N/n_c$ unique uncontrollable factor settings. This implies that the simulation time for a crossed design with the same number of unique uncontrollable factor settings becomes n_c times longer. For a fixed experimental design size, the proposed design optimally covers the uncontrollable factor space in terms of overall uniformity. In the modeling and analysis of the simulation output, the uniformity provides good coverage for the uncontrollable factor effects.

REFERENCES

Fang, K.-T., R. Li, and A. Sudjianto. 2006. *Design and modeling for computer experiments*. Chapman & Hall/CRC.

Fang, K.-T., D. K. J. Lin, and M.-Q. Liu. 2003. Optimal mixed-level supersaturated design. *Metrika* 58 (3): 279–291.

Fang, K.-T., and C.-X. Ma. 2001. Wrap-around l2-discrepancy of random sampling, latin hypercube and uniform designs. *Journal of Complexity* 17 (4): 608–624.

Hickernell, F. 1998a. *Random and quasi-random point sets*, Chapter Lattice rules: How well do they measure up?, 106–166. Springer-Verlag, New York.

Hickernell, F. J. 1998b. A generalized discrepancy and quadrature error bound. *Mathematics of Computation* 67 (221): 299–322.

Kleijnen, J. P. 2008. *Design and analysis of simulation experiments*. Springer.

Krahl, D. 2002. The extend simulation environment. In *Proceedings of the 2002 Winter Simulation Conference*, 205–213.

Montgomery, D. C. 2005. *Design and analysis of experiments*. 6th ed. John Wiley and Sons, Inc.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn. 1989. Design and analysis of computer experiments. *Statistical Science* 4 (4): 409–423.

Sanchez, S. M. 2000. Robust design: Seeking the best of all possible worlds. In *Proceedings of the 2000 Winter Simulation Conference*, 69–76.

Santner, T. J., B. J. Williams, and W. I. Notz. 2003. *The design and analysis of computer experiments*. Springer.

Taguchi, G. 1987. *System of experimental design, volumes 1 and 2*. UNIPUB/Krauss International, White Plains, New York.

AUTHOR BIOGRAPHIES

CHRISTIAN DEHLENDORFF is a Ph.D. student at the Department of Informatics and Mathematical Modelling, Technical University of Denmark. His email and web addresses are <cd@imm.dtu.dk> and <http://www.imm.dtu.dk/~cd>.

MURAT KULAHCI is an Associate Professor at the Department of Informatics and Mathematical Modelling, Technical University of Denmark. His email address is <mk@imm.dtu.dk>.

KLAUS KAAE ANDERSEN is an Associate Professor at the Department of Informatics and Mathematical Modelling, Technical University of Denmark. His email address is <kka@imm.dtu.dk>.