Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

THE VARIABILITY IN DESIGN-QUALITY MEASURES FOR MULTIPLE TYPES OF SPACE-FILLING DESIGNS CREATED BY LEADING SOFTWARE PACKAGES

Thomas W. Lucas

Department of Operations Research
Naval Postgraduate School
1411 Cunningham Road
Monterey, CA 93943, USA

Jeffrey D. Parker

Combat Development & Integration
U.S. Marine Corps
3300 Russell Road
Quantico, VA 22134, USA

ABSTRACT

Space-filling designs (SFDs) underpin many large-scale simulation studies. The algorithms that construct SFDs are mostly stochastic and cannot guarantee that optimal solutions can be found within a practical amount of time. This paper uses massive experimentation to find the empirical distributions of a diverse set of design-quality measures in highly-used classes of SFDs constructed by leading software packages. The objective is to provide simulation practitioners with a better understanding of what they can expect from different SFD choices. The results show substantial variability in measures of correlation and space-fillingness in the design classes and dimensions investigated. Therefore, computer experimenters should generate and assess several candidate designs using different random-number-generator seeds to reduce the risk of using a poor design simply due to random chance. We also find that in the largest designs investigated, the uniform designs generally perform best for both our correlation and uniformity measures.

1 INTRODUCTION

The simulations researchers use often contain many input variables, nonlinear relationships, stochastic (or pseudo-random) elements, and generate multiple diverse responses (Kleijnen et al. 2005). To obtain insight from such complex computer models, researchers often construct metamodels of the responses as functions of the input variables (Law and Kelton 2000). The metamodels that can be fit, and hence, the insights that can be gleaned, depend critically on the design of experiments (DOE) used to construct the metamodels. The DOE specifies the settings of the input factors over the simulation runs.

Computer experimenters desire designs that "allow one to fit a variety of models and should provide information about all portions of the experimental region" (Santner et al. 2018, p. 148). They also seek designs with zero or minimal correlations among columns in the design matrix since multicollinearity adversely effects many statistical techniques (Montgomery 2013). Achieving these design objectives is challenging when there are many input factors and limits on the number of experiments. In such situations, computer experimenters often use *space-filling designs* (SFDs), see Sanchez et al. (2020). Intuitively, an SFD has "points everywhere in the experimental region with as few gaps or holes as possible" (Joseph 2016, p. 29). Or, as Lin and Tang (2015, p. 593) write, SFDs "fill a bounded design region as uniformly as possible."

There are many classes of SFDs and methods to construct them (Fang et al. 2005). Unfortunately, building large-scale SFDs is usually challenging. Most construction algorithms use stochastic heuristic search methods that do not guarantee an optimal design will be found within a reasonable amount of time (Lin and Tang 2015). Depending on the random-number-generator seed, initial conditions for the search, stopping criteria, processing time, computing power, and search-algorithm parameters, different "optimal" solutions with substantially disparate measures of design quality can be obtained (Parker 2022).

From among the many available classes of SFDs, what should simulation experimenters use, and when? What design-quality measures can they expect? What risk are they taking from the randomness in most design generation algorithms? Of course, the answers depend on the experimenter's goals and will vary greatly in different circumstances. To provide insight to simulation researchers about their design options, this investigation moves toward addressing Jin et al.'s (2003, p. 554) call for a "thorough future investigation" of the "many optimality criteria available in the literature" for SFDs. Two decades later, we cannot find substantial progress towards understanding the performance, variability, and relationships among different measures of design quality for the most used classes of SFDs. This research takes a long-overdue first step in this direction.

This paper presents the empirical distributions of three diverse design-quality measures in five commonly used classes of SFDs built using leading software packages. Section 2 specifies the experimental setting, defines our notation, explains the three design-quality measures we investigate, and introduces the five classes of SFDs explored. Section 3 uses a sequence of box plots to display the empirical distributions of a measure of correlation, a distance-based space-filling criterion, and a uniformity measure for nine design sizes in the five classes of SFDs. Section 4 summarizes the main findings and identifies areas for future research.

2 BACKGROUND

2.1 The Simulation Setting

We consider situations where simulation experimenters select continuous input values over a rectangular region. The $n \times k$ *design matrix*, **X**, specifies the simulation's input settings for *n* runs involving *k* continuous factors (i.e., input variables). Row *i* of **X**, denoted x_i, is called a *design point* (DP). DP x_i specifies the values for each of the *k* factors for which the simulation will be run, for $i = 1, \ldots, n$. The c^{th} column of **X**, which we label as X_c, stipulates the settings for factor *c* over the *n* DPs. x_{ic} specifies the input value for factor *c* in DP *i*. Furthermore, to make our design comparisons are meaningful, we scale the input so that $x_i \in [0,1]^k = \chi$, the *k*-dimensional unit cube that comprises the experimental region.

2.2 Design-Quality Measures

Practitioners use a variety of measures to assess their designs (Santner et al. 2018). In situations involving considerable *a priori* uncertainty about the forms of many diverse responses, these measures are typically based on **X** and mostly fall into three broad classes, which are measures of correlation, uniformity, and distance. This subsection defines the three prominent design-quality measures in the literature we study, one from each category.

2.2.1 A Measure of Correlation

Minimizing the correlation among factors in a design has been a goal since the early days of the science of DOE (Fisher 1925). Following Moon et al. (2012, p. 378), we prefer designs with zero or minimal correlations among columns in **X** "to allow independent [or nearly independent] assessments of the [main] effects of the different inputs." The correlation between columns X_c and X_d of **X** is given by:

$$\rho_{c,d} = \frac{\sum_{i=1}^{n}\left[(x_{ic} - \bar{X}_c)(x_{id} - \bar{X}_d)\right]}{\sqrt{\left[\sum_{i=1}^{n}(x_{ic} - \bar{X}_c)^2\right]\left[\sum_{i=1}^{n}(x_{id} - \bar{X}_d)^2\right]}}.$$

Here, $\bar{X}_c$ and $\bar{X}_d$ are the means of columns X_c and X_d, respectively.

We express the maximum absolute pairwise (*map*) correlation among columns of **X** as

$$\rho_{map} = \max\left\{|\,\rho_{c,d}\,|,\ \forall\ c \neq d\right\}.$$

ρ_{map} quantifies how far from orthogonal the columns of **X** are, with $\rho_{map} = 0$ indicating **X** is orthogonal. Minimizing ρ_{map} bounds the worst-case pairwise correlation among columns in design matrix **X**. A design with $\rho_{map} \leq 0.05$ is called *nearly orthogonal* (Hernandez et al. 2012a). Designs with lower ρ_{map} values are preferred when evaluated by this criterion.

2.2.2 A Measure of Distance

Distance measures are likely the most common approach to constructing and quantifying the space-fillingness of a design (Joseph 2016). For a fixed number of DPs in χ, large inter-point distances are desirable; i.e., the points are spread out rather than many being clustered in a portion of a bounded space. The p^{th} order distance between any two design points x_i and x_j is defined as

$$d_p(x_i, x_j) = \left(\sum\nolimits_{c=1}^{k} | x_{ic} - x_{jc} |^p\right)^{1/p},$$

where $p = 1$ is the Manhattan distance and $p = 2$ is the Euclidean distance.

An approach to measuring the largest gap or hole in χ is by the maximum distance from any point x in χ from its nearest DP. A common design goal is to find the design that minimizes this “worst case” distance for all possible designs of the same dimension (i.e., n and k). A design that achieves this is called a *minimax* (mM) *distance design* (Johnson et al. 1990). Since finding minimax distance designs is computationally challenging, *maximin* (Mm) *distance designs* (Johnson et al. 1990) are more common in practice. A design is called an *Mm distance design* if it maximizes the minimum distance between any two DPs, with larger values preferred. However, finding an Mm distance design is still difficult for large designs. Morris and Mitchell (1995) developed Mm distance designs with **X** constrained to a Latin hypercube design (McKay et al. 1979) with evenly spaced levels for each factor. This ensures good projective properties in each factor’s subspace. These designs are known as *maximin Latin hypercube designs* (MmLHDs). According to Joseph (2016, p. 31) “MmLHD seems to be the most commonly used experimental design for computer experiments in practice because of its simplicity and availability in software packages.” While MmLHDs are common in practice, Joseph (2016) found that minimizing the *maximum projection* (MaxPro) criterion leads to better space-fillingness in projections of **X** into low-dimensional subspaces. This feature is highly-valued when only a small number of factors have significant effects on responses.

2.2.3 A Measure of Uniformity

Another approach to measuring the space-fillingness of a design is to quantify how uniformly the DPs are spread throughout χ (Fang 1980). Specifically, the *uniformity* or *discrepancy* of design **X** is defined as

$$D^*(\mathbf{X}) = \max_{z \in \chi} \left| \frac{N(\mathbf{X}, R_z)}{n} - vol(R_z) \right|.$$

Here, R_z is the subregion $[0, z_1) \times \ldots \times [0, z_k)$ within χ, $vol(R_z)$ is the volume of subregion R_z, and $N(\mathbf{X}, R_z)$ is the number of design points of **X** within R_z. The objective is to construct **X** with $D^*(\mathbf{X})$ as close to zero as possible. Since $D^*(\mathbf{X})$ is computationally burdensome, following Hickernell (1998), this research uses the *modified L_2 discrepancy* $(ML_2)^2$ to quantify the discrepancy of a design:

$$\left(ML_2\right)^2 = \left(\frac{4}{3}\right)^k - \frac{2^{1-k}}{n} \sum_{i=1}^{n} \prod_{c=1}^{k} \left(3 - x_{ic}^2\right) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \prod_{c=1}^{k} \left[2 - \max(x_{ic}, x_{jc})\right].$$

2.3 SFD Construction Algorithms, Design Classes, and Generating Software

Constructing large-scale designs that optimize space-filling measures is challenging due to nonlinear objective functions, integer constraints, and high dimensionality. Most SFD construction methods use heuristic stochastic search algorithms, such as simulated annealing or genetic algorithms, which cannot guarantee that an optimal design will be found within a constrained time limit (Lin and Tang 2015). Indeed, Parker (2022) found that variability remained even when some search algorithms ran for up to two days on a MacBook Pro with a 3.1GHz processor, Intel Core i7, and a memory of 16GB. For all stochastic optimization methods, each random instantiation will likely provide a different "optimal" solution, thereby generating a distribution of design-quality measures.

This paper explores the variability in design-quality measures for five diverse design classes and software at their default settings. From the many possibilities, we chose to investigate SFD classes that are featured in the leading texts on computer experiments (e.g., Santner et al. 2018; Fang et al. 2005), among the most frequently used in our experience, available to most practitioners, or reported to be among the most used (Joseph 2016; Wang et al. 2020). Using the "space-filling designer" in JMP (SAS 2021), sphere-packing (i.e., Mm distance) designs and maximin distance Latin hypercube designs (MmLHDs) were constructed. We generated state-of-the-art MaxPro designs and uniform designs (UDs) using the R software packages *MaxPro* (Ba and Joseph 2018) and *UniDOE* (Zhang et al. 2018). Following the recommendation of Ba and Joseph (2018), the MaxPro designs were initialized with a MaxProLHD (i.e., first optimizing the MaxPro criterion with design matrix **X** constrained to be an LHD) and then minimized with the LHD restriction removed. Custom R functions were used to generate the LHDs.

3 THE EMPIRICAL DISTRIBUTIONS OF THREE DIVERSE DESIGN-QUALITY MEASURES FOR FIVE SFD CLASSES OF A VARIETY OF SIZES

This section shows the empirical distributions of three design-quality measures in five SFD classes for nine design sizes (i.e., n and k combinations). These results provide guidance to simulation experimenters in choosing the design class to use and what quality measures they can expect.

3.1 Our Experiments

Our experiments vary three factors. The design class has five categorical levels (LHD, MaxPro, MmLHD, SphereP, and UniDOE). We vary k and n to yield nine design dimensions, with k = 5, 10, and 20 and $n = k+1$, $3k+2$, and $10k$. This allows us to explore the distributions of the measures in each class for three levels of k and three levels relating to design density (i.e., the number of DPs per input variable), which we classify as low, mid, and high. A full-factorial design was run, yielding 45 DPs. For each combination, we generated 100 independent designs using JMP or R at their default settings—yielding 4,500 designs. For each of these designs, our design-quality measures were calculated. Thus, all the measures were calculated from the same designs. When $n = k+1$, the low-density designs are fully saturated. Saturated designs are more common when the number of feasible simulation runs is constrained. In cases where $n = 10k$, these high-density designs have more degrees of freedom for fitting metamodels. Having $n = 10k$ matches the rule-of-thumb guidance given by Loeppky et al. (2009). Designs with $n = 3k+2$, our mid-density designs, have DP densities between the two extremes. This mid-level value was chosen since nearly orthogonal and good space-filling Latin hypercubes are known to exist at this design size (Cioppa and Lucas 2007).

3.2 Distributions of ρ_{map}

We begin our exploration by looking at the distributions of ρ_{map} in the five SFD classes for our smallest design size. Figure 1 displays side-by-side box plots of ρ_{map} for $n = 6$ and $k = 5$ (i.e., 6×5 design matrices). Each box plot shows the empirical distributions from 100 independent randomly generated designs. Since there are five design types, the figure encompasses 500 ρ_{map} values. The ρ_{map} values across the designs range from 0.030 (i.e., a nearly orthogonal sphere-packing design) to 1.000 (i.e., there is a completely confounded LHD). For the LHD, MaxPro, and UniDOE designs, the ρ_{map} values have substantial variability, with ranges of around 0.50. Thus, users of these designs, at this size, face considerable risk if they generate a single "optimal" design.

For these small fully-saturated designs, sphere-packing designs consistently provide the lowest ρ_{map} values and have the least variability—that is, they would be preferred under this criterion. Only three of the 500 designs are nearly orthogonal, and all three are sphere-packing designs. The intuition is that JMP's sphere-packing algorithm optimizes on the Mm Euclidean distance and low ρ_{map} values can be a byproduct of placing DPs near the corners of χ, as in fractional factorial designs. The median ρ_{map} value for sphere-packing designs is 0.092. The next three lowest median ρ_{map} values are from UniDOE, MaxPro, and MmLHDs, with 0.200, 0.221, and 0.257, respectively—which are roughly comparable. Interestingly, the ordering of the 3rd quartile for those three classes is reversed. While LHDs are frequently used in practice, due in part because they are easy to construct, the highest ρ_{map} values occur in LHDs, by a considerable margin, with a median ρ_{map} value of 0.771. Thus, there is likely benefit in using more sophisticated algorithms to construct SFDs than LHDs, such as the other four types. A more expansive treatment of ρ_{map} values in LHDs is contained in Hernandez et al. (2012b). The ranges in ρ_{map} values obtained within each design class highlight the importance of practitioners generating and evaluating multiple candidate designs.

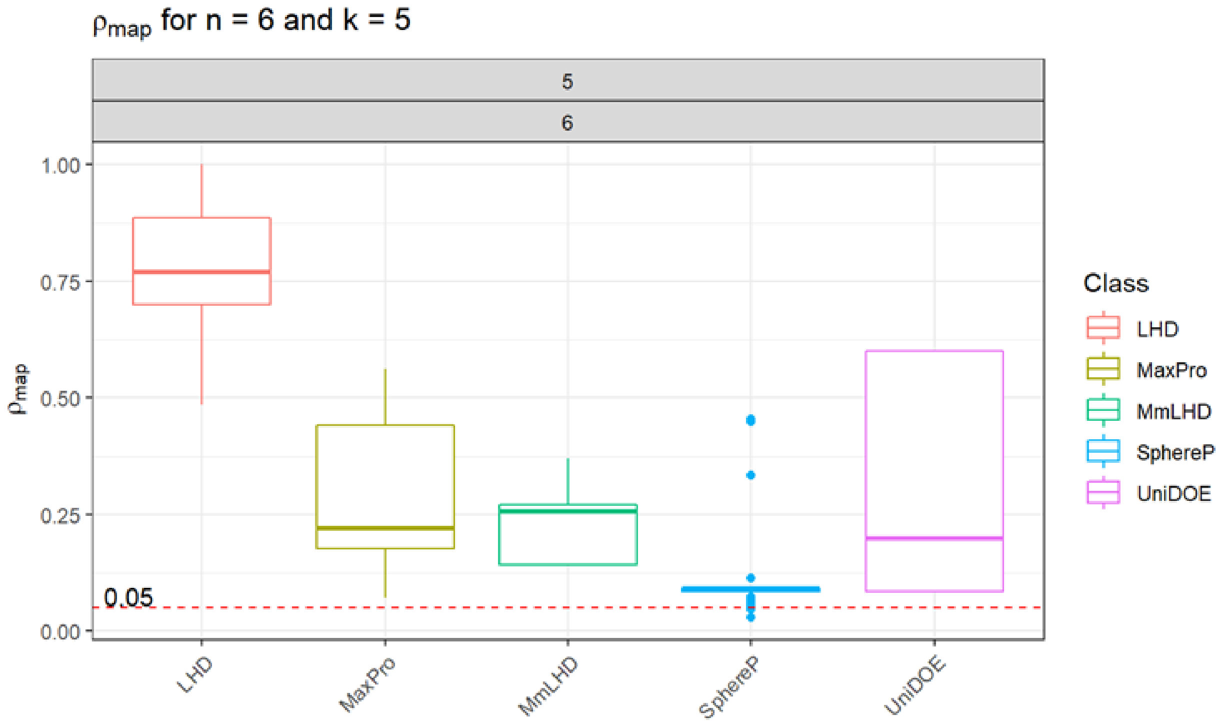

Figure 1: Box plots of ρ_{map} for 100 6×5 LHD, MaxPro, MmLHD, SphereP, and UniDOE designs.

We extend our exploration by increasing the design density, while holding k at five. Figure 2 shows the distributions of ρ_{map} in our five SFD classes for k = 5 and n = 6, 17, and 50. Figure 1 is the leftmost panel in this graphic. As n increases, i.e., greater design density, ρ_{map} values generally trend lower, as does their variance. However, there is a notable exception, as 17×5 sphere-packing designs have generally higher ρ_{map} values than 6×5 sphere-packing designs. We also see that the preferred design class by this measure changes as n increases. For n = 50, our highest density designs, the R software package *UniDOE* designs perform best with respect to ρ_{map}, with a median ρ_{map} value of 0.047 and most of its designs being nearly orthogonal. The next best performing design classes when n = 50 and k = 5, with respect to median ρ_{map}, are, in order, sphere-packing (0.080), MmLHD (0.096), MaxPro (0.124), and LHD (0.253). When n = 6 or 17, the sphere-packing designs have the lowest (i.e., best) ρ_{map} values. For all design densities when k = 5, LHDs perform worst. Interestingly, the third-best median ρ_{map} value switches from MaxPro to MmLHD between design sizes 17×5 and 50×5.

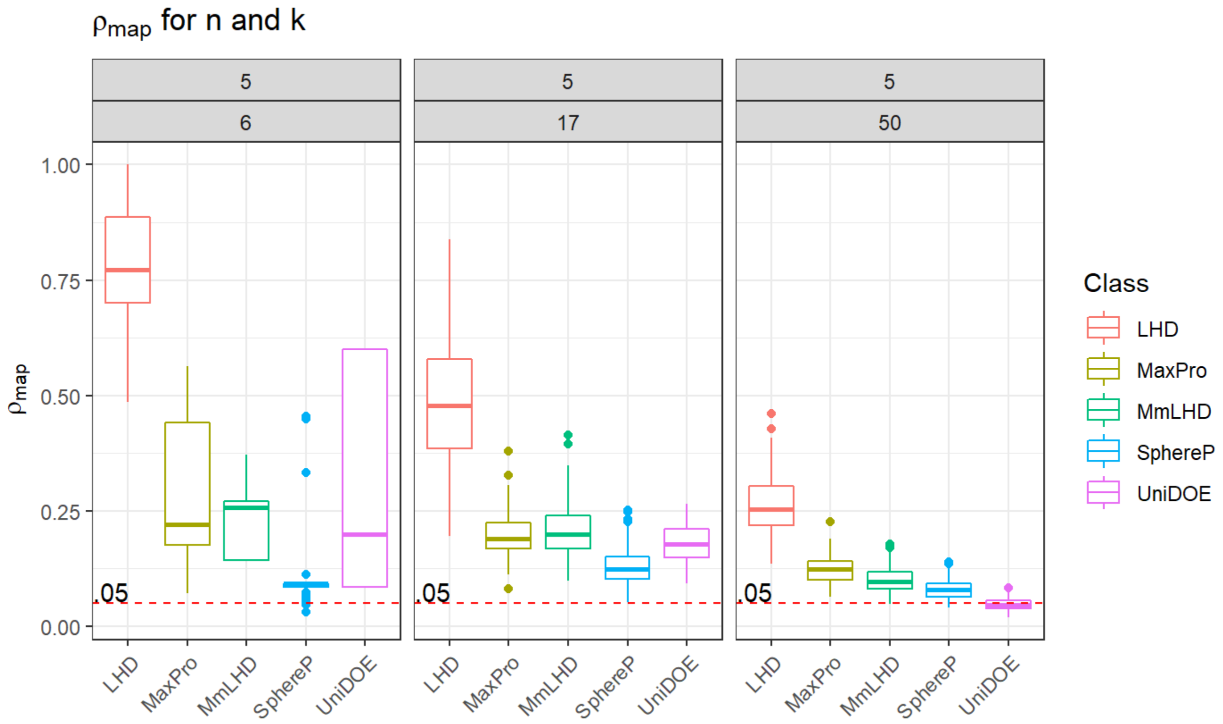

Figure 2: Box plots of ρ_{map} for LHD, MaxPro, MmLHD, SphereP, and UniDOE designs when k = 5.

Our investigation continues by looking at designs with more factors. Figure 3 appends two rows (for k = 10 and k = 20) to what is displayed in Figure 2, forming nine total panels of side-by-side box plots. Panels in the same row show box plots for designs with the same number of factors. Panels in the same column display box plots for designs with similar DP densities, i.e., with $n/k = 1+1/k$, $3+2/k$, and 10, for columns one through three, respectively—i.e., our low, mid, and high categories of design density. Smaller designs are towards the top left and larger designs are towards the bottom right. These panels cover the nine combinations we explore of the number of factors and the three design density categories.

Looking across the nine panels in Figure 3, ρ_{map} values range from 0.020 to 1.000, with an overall median value of 0.214. Except for 6×5 sphere-packing designs, ρ_{map} values for saturated designs (column one) are consistently among the highest obtained. As design density increases, i.e., moving left to right within a row, ρ_{map} values and their range tend to decrease—with the notable exception of sphere-packing designs for k = 5 when n goes from 6 to 17. When $n = 10k$, all 1,500 designs have ρ_{map} values less than 0.50. For large designs with mid to high density, those towards the lower right of the figure, we observe that the *UniDOE* R software package consistently generates designs with the lowest ρ_{map} values and with the least variability—and are thus preferred by this measure. Moreover, the high density UniDOE designs are often nearly orthogonal. No other SFD class of any size produces more than an occasional nearly orthogonal design. Within all nine panels, the highest ρ_{map} values occur in LHDs. A clear ordering for design class preference using ρ_{map} emerges in the four lower-right panels, which correspond to the larger designs. The best-performing class by this measure is UniDOE, followed by sphere packing. The MmLHD and MaxPro designs seem roughly comparable, while the LHD designs perform the worst. It's worth noting that none of these design-generation algorithms explicitly consider ρ_{map}.

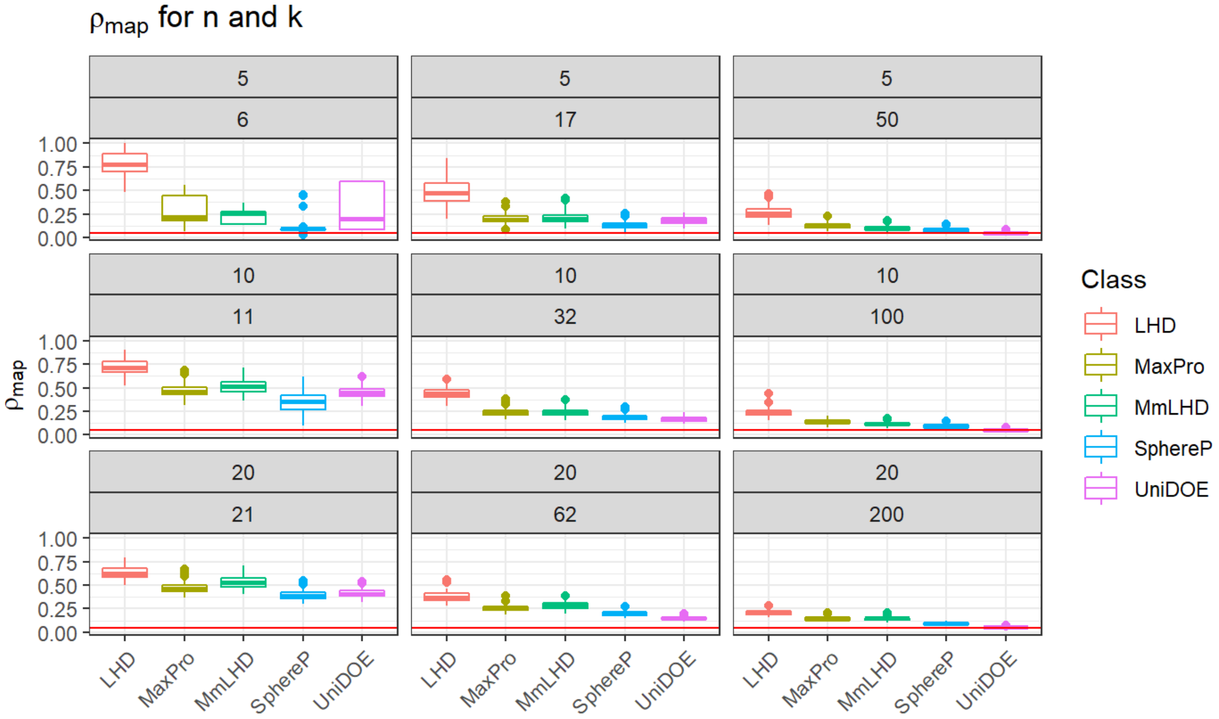

Figure 3: Box plots of ρ_{map} for LHD, MaxPro, MmLHD, SphereP, and UniDOE designs.

Since we are most interested in simulations with many factors, we extract and display the bottom row (i.e., when $k = 20$) of Figure 3. This "zooming in" allows us to better identify differences obscured by plot size. Figure 4 shows the box plots of ρ_{map} for each design type for designs of sizes 21×20, 62×20, and 200×20. We see that in the large, saturated designs (left panel), all ρ_{map} values are greater than 0.30. Notably, nearly orthogonal LHDs of this dimension have been constructed by Hernandez et al. (2012a). As n increases, ρ_{map} values trend lower for all design classes and ρ_{map} variability decreases. In fact, the maximum (i.e., worst) ρ_{map} value from all of the 500 200×20 designs in the far-right panel is 0.280, which is lower than (i.e., preferred to) the minimum ρ_{map} value (0.300) from all 500 21×20 designs in the left panel. This highlights how valuable larger sample sizes (i.e., bigger n) can be to experimenters. We also see that UniDOE designs perform substantially better than the other methods for our largest (i.e., 200×20) designs, with many nearly orthogonal and a median ρ_{map} value of 0.052. In fact, the largest ρ_{map} value in the 200×20 UniDOE designs is 0.075. This is lower than the best ρ_{map} value for the LHD, MaxPro, and MmLHD designs. The sphere-packing designs are a clear number two in the 62×20 and 200×20 designs. For these design dimensions, MaxPro and MmLHD are generally comparable, while LHDs consistently yield the poorest-performing designs according to the ρ_{map} criterion.

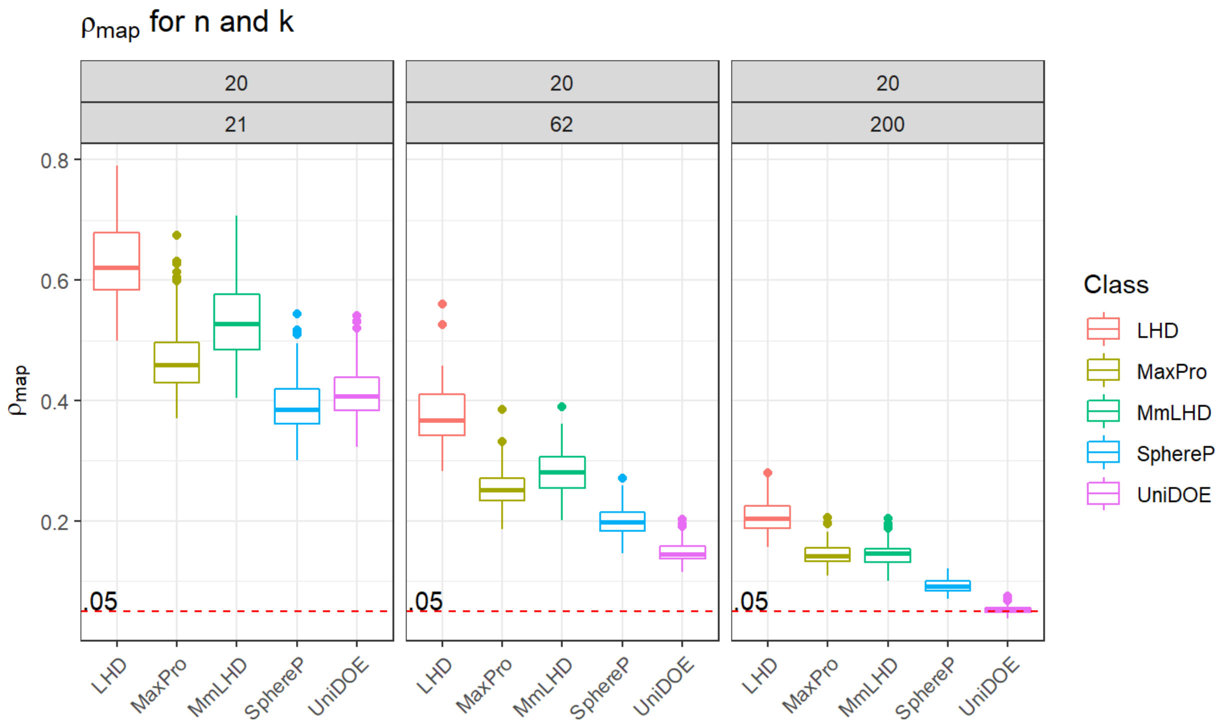

Figure 4: Box plots of ρ_{map} for LHD, MaxPro, MmLHD, SphereP, and UniDOE designs when k = 20.

3.3 Distributions of Mm Euclidean Distance

Mm distance criteria are widely used to construct and assess SFDs (Joseph 2016). Figure 5 presents comparative box plots for the Euclidean Mm distance across nine design dimensions. Greater Mm distances (higher values in the plots) are generally preferred for a given design size and bounded region, meaning no two DPs are “closer than they need to be.” As expected, the sphere-packing designs, which optimize over this measure without any constraints, other than requiring all DPs be in χ, perform best for all design sizes. In fact, in all panels, the worst-performing sphere-packing design is greater than the best of any other design. The sphere-packing design’s dominance by this measure grows as k increases. The MmLHDs tend to have the second-highest average Mm distance, except for the saturated designs with k = 10 and 20, where the MaxPro designs beat them. The LHDs consistently have the lowest (i.e., worst) Mm distances, even though LHDs are constructed to have optimal space-filling in projections into each factor’s subspace. For higher design densities, especially when $n = 10k$, the MmLHDs emerge as the second-best option by this measure.

We have seen that sphere-packing designs perform best with respect to Mm distance for all design sizes explored, often by a considerable amount. Thus, to obtain more discrimination in examining the other design classes in our largest designs, we “zoom in” on the lower right panel with the pest-performing sphere-packing designs removed, see Figure 6. We see that MmLHDs dominate MaxPro designs with respect to this measure. The MaxPro designs appear slightly better, as a whole, than the UniDOE designs. Again, the LHDs have the least desired (i.e., lowest) Mm values. We also observe much less variability in Mm distance for MmLHDs than in the other SFD classes in these large designs.

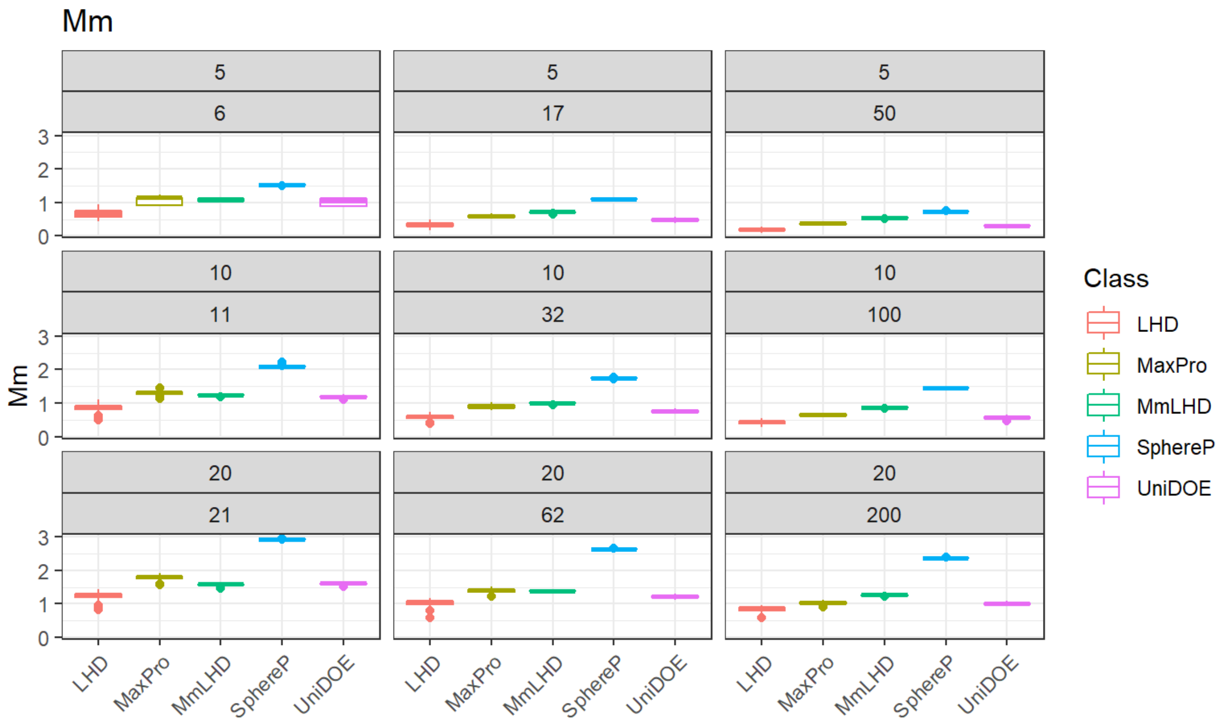

Figure 5: Box plots of Mm distances for LHD, MaxPro, MmLHD, SphereP, and UniDOE designs.

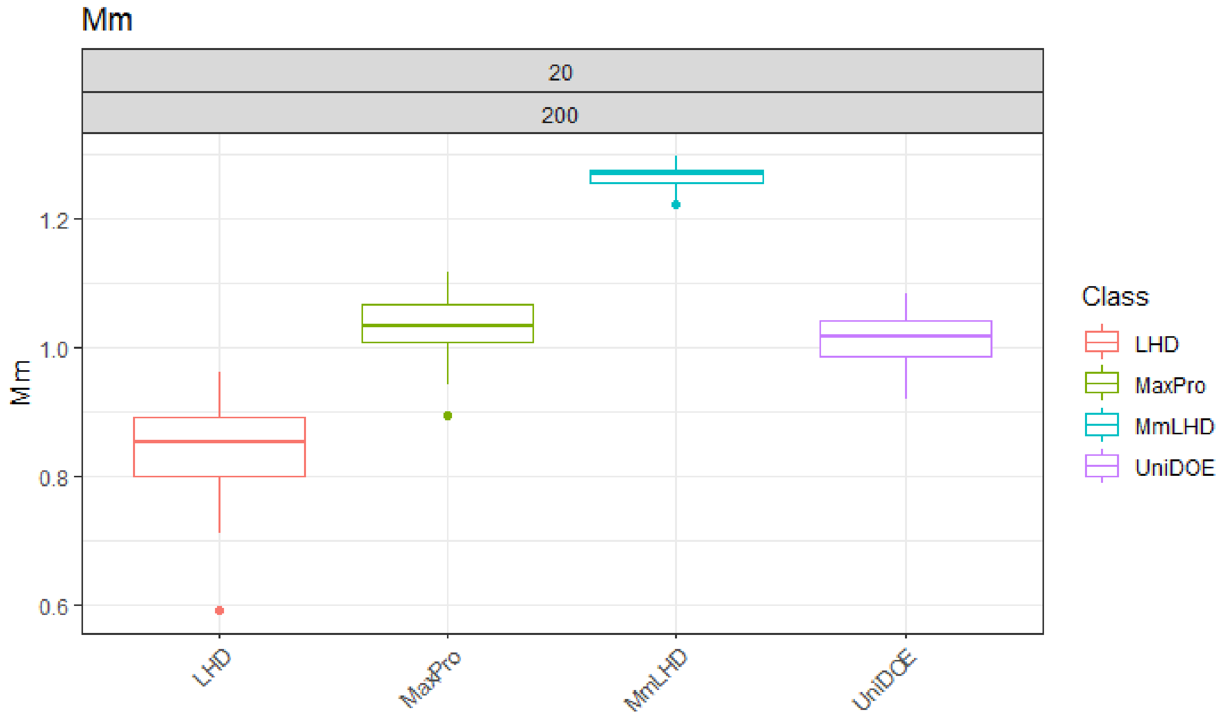

Figure 6: Box plots of Mm distances for 200×20 LHD, MaxPro, MmLHD, and UniDOE designs.

3.4 Distributions of $(ML_2)^2$

The final measure we explore is our uniformity measure, a very different measure of space-fillingness than the distance-based ones. $(ML_2)^2$ is a surrogate measure of the discrepancy between the empirical cumulative distribution function (CDF) from the design's DPs compared to the CDF of a theoretical k-dimensional uniform distribution over χ. Figure 7 presents nine panels for different design dimensions of side-by-side box plots for $(ML_2)^2$. Again, read rowwise for k = 5, 10, and 20 and columnwise for n = k+1, 3k+2, and 10k. As expected, the uniDOE designs, which optimize for discrepancy, have the lowest median $(ML_2)^2$ values in each panel, and are thus preferred by this measure. This is difficult to see in the figure due to the great differences in ranges, which obscure the results in many subpanels. Optimizing for the Mm distance criterion in the sphere-packing designs results in the highest (i.e., worse) discrepancies, by far, for all design sizes. The insight is that spreading-out DPs using the Mm distance criterion likely increases the number of DPs near the boundary of χ, resulting in a nonuniform distribution of points in the interior. This effect is most apparent when k = 20. The difference in $(ML_2)^2$ values between the sphere-packing designs and MmLHDs shows a benefit of imposing an LHD structure. Also, note how much worse the $(ML_2)^2$ values are in the 21×20 designs, i.e., the large-saturated designs. This pattern arises because it is far more challenging to have DPs uniformly cover χ when n is small and k is large.

Other insights gleaned from the figure include that $(ML_2)^2$ values decrease rapidly as design density (i.e., n) increases, though proportionally less so for sphere-packing designs. For the designs that are ranked between uniDOE and sphere-packing, the preference by median $(ML_2)^2$ values is MmLHD, LHD, and MaxPro, respectively, in eight of the nine panels. The exception is for our 200×20 designs, where the preference ordering is MaxPro, MmLHD, and LHD.

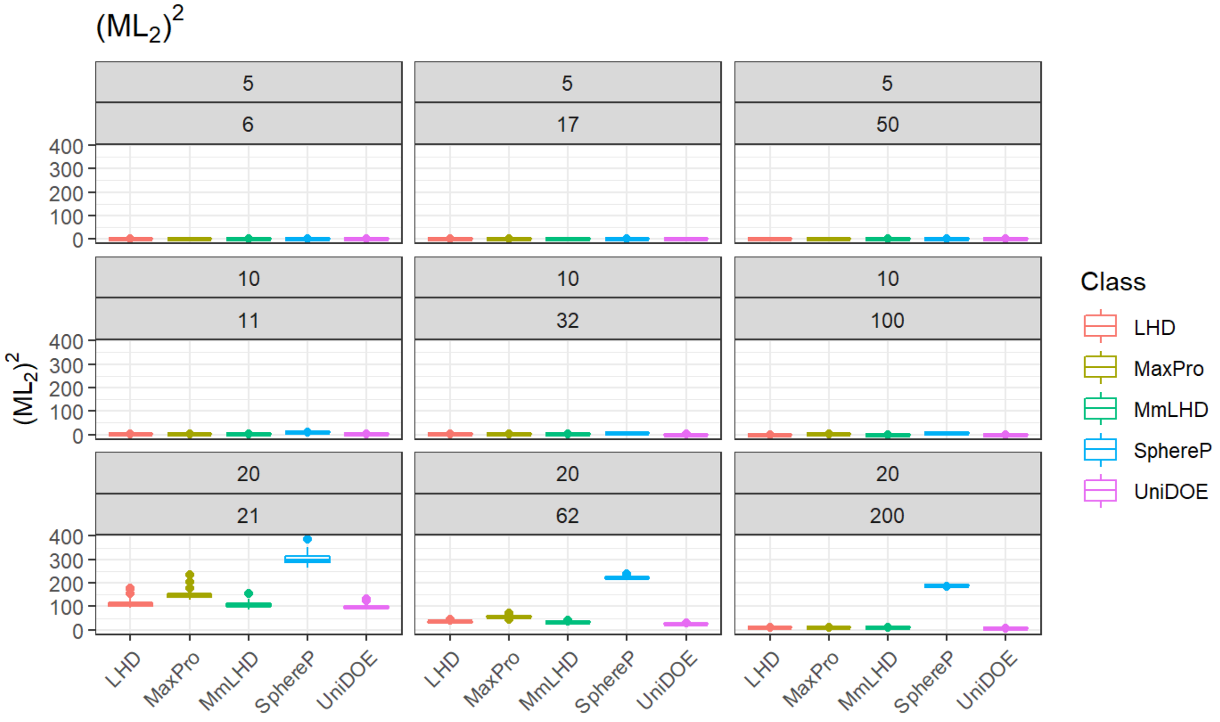

Figure 7: Box plots of $(ML_2)^2$ values for LHD, MaxPro, MmLHD, SphereP, and UniDOE designs.

4 CONCLUSION

Complex computer simulations are increasingly being used by scientists, businesses, and governments (Powers et al. 2012). Modern SFDs help us efficiently learn from simulation experiments. Unfortunately, as Jin et al. (2003) noted, there is a dearth of understanding of the many measures of design quality for SFDs. The results in this paper can assist simulators in selecting their design and the quality they can expect. First and foremost, we show there is substantial variability in measures of correlation and space-fillingness in most of the design classes and dimensions investigated. It follows that practitioners should generate and assess several candidate designs using different random-number-generator seeds to reduce the risk of using a poor design due simply to random chance. As design density increases, the design-quality measures tend to steadily improve and their variability decreases. This highlights and quantifies how larger sample sizes benefit experimenters. While the recommended design depends on the quality measure and design size, for designs with many factors and runs, the uniform designs stood out as clearly the best for our correlation and uniformity measures.

Other top-level lessons learned from our experiments include the following. With respect to ρ_{map}, for our small, saturated designs, sphere-packing designs perform best by a substantial margin. For high-density designs, the uniDOE designs are the best and often nearly orthogonal. In all design dimensions, LHDs perform worst. For Mm Euclidean distance, sphere-packing designs perform best for all design sizes explored, often by a considerable amount, especially for large k. For higher design densities, especially when $n = 10k$, MmLHDs emerge as the second-best option. Regarding $(ML_2)^2$, in all nine design dimensions, the uniDOE designs have the lowest (i.e., best) median values and sphere-packing designs have the highest (i.e., worst) median values. Results from more experiments, measures, and design classes, as well as additional insights, are available in Parker (2022).

While this exploration is far more comprehensive than any other of this type in the literature the authors are aware of, it barely scratches the surface of what is possible. To begin with, several other measures, design classes, design dimensions, software packages, and search algorithm settings can be investigated. In addition, all our designs were created before any simulation experiments were run. There is a rich literature on sequential designs for computer experiments (Kleijnen 2015). The above type of analysis can be extended to sequential designs. Of course, with many sequential procedures, the results may depend, i.e., adapt, dramatically based on responses. In such cases, some response forms would need to be assumed. Finally, our analysis focuses on univariate distributions of design-quality measures. It is also of interest to quantify the correlations and other relationships among the measures.

ACKNOWLEDGMENTS

This research was supported by a grant from the Operational Analysis Division (OAD) of the United States Marine Corps. The authors thank W. David Kelton and the referees for helping improve our paper.

REFERENCES

Ba, S., and V. R. Joseph. 2018. *MaxPro: Maximum Projection Designs (CRAN).* https://CRAN.R-project.org/package=MaxPro, accessed May 9th, 2022.

Cioppa, T. M., and T. W. Lucas. 2007. “Efficient Nearly Orthogonal and Space-filling Latin Hypercubes”. *Technometrics* 49(1):45–55.

Fang, K. T., R. Li, and A. Sudjianto. 2005. *Design and Modeling for Computer Experiments*. CRC press.

Fang, K. T. 1980. “Uniform Design: Application of Number-theoretic Methods in Experimental Design.” *Acta Mathematicae Applicatae Sinica* 3:363–372.

Fisher R. A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Hernandez, A. S., T. W. Lucas, and M. Carlyle. 2012a. “Constructing Nearly Orthogonal Latin Hypercubes for any Nonsaturated Run-variable Combination”. *ACM Transactions on Modeling and Computer Simulation* 22(4):1–17.

Hernandez, A. S., T. W. Lucas, and P. J. Sanchez. 2012b. “Selecting Random Latin Hypercube Dimensions and Designs Through Estimation of Maximum Absolute Pairwise Correlation”. In *Proceedings of the 2012 Winter Simulation Conference,* edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, Article 25. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Hickernell, F. J. 1998. "A Generalized Discrepancy and Quadrature Error Bound". *Mathematics of Computation* 67(221):299–322.

Jin, R., W. Chen, and A. Sudjianto. 2003. "An Efficient Algorithm for Constructing Optimal Design of Computer Experiments". In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 37009:545-554.

Johnson, M. E., L. M. Moore, and D. Ylvisaker. 1990. "Minimax and Maximin Distance Designs". *Journal of Statistical Planning and Inference* 26(2):131–148.

Joseph, V. R. 2016. "Space-filling Designs for Computer Experiments: A Review". *Quality Engineering* 28(1):28–35.

Kleijnen, J. P. C. 2015. *Design and Analysis of Simulation Experiments*. 2nd ed. Springer International Publishing.

Kleijnen, J. P. C., S. M. Sanchez, T. W. Lucas, and T. M. Cioppa. 2005. "State-of-the-art Review: A User's Guide to the Brave New World of Designing Simulation Experiments". *INFORMS Journal on Computing* 17(3):263–289.

Law, A. M., and W. D. Kelton. 2000. *Simulation Modeling and Analysis*. 3rd ed. New York: McGraw-Hill, Inc.

Lin, C. D., and B. Tang. 2015. "Latin Hypercubes and Space-filling Designs". In *Handbook of Design and Analysis of Experiments*, edited by A. M. Dean, M. Morris, J. Stufken, and D. Bingham, 593–625. Boca Raton: CRC Press.

Loeppky, J. L., J. Sacks, and W. J. Welch. 2009. "Choosing the Sample Size of a Computer Experiment: A Practical Guide". *Technometrics* 51(4):366–376.

McKay, M. D., R. J. Beckman, and W. J. Conover. 1979. "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code". *Technometrics* 21(2):239–245.

Moon, H., A. M. Dean, and T. J. Santner. 2012. "Two-stage Sensitivity-based Group Screening in Computer Experiments". *Technometrics* 54(4):376–387.

Morris, M. D., and T. J. Mitchell. 1995. "Exploratory Designs for Computational Experiments". *Journal of Statistical Planning and Inference* 43(3):381–402.

Montgomery, D. C. 2013. *Design and Analysis of Experiments*. 8th ed. Hoboken, NJ: John Wiley & Sons, Inc.

Parker, J. D. 2022. *Extending and Improving Designs for Large-Scale Computer Experiments*. Ph.D. thesis, Operations Research Department, Naval Postgraduate School, Monterey, California. https://calhoun.nps.edu/handle/10945/70717, accessed 29th March 2023.

Powers, M. J., S. M. Sanchez, and T. W. Lucas. 2012. "The Exponential Expansion of Simulation in Research". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, Article 138. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sanchez, S. M., P. J. Sanchez, and H. Wan. 2020. "Work Smarter, Not Harder: A Tutorial on Designing and Conducting Simulation Experiments". In *Proceedings of the 2020 Winter Simulation Conference*, edited by K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 1128–1142. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Santner, T. J., B. J. Williams, and W. I. Notz. 2018. *The Design and Analysis of Computer Experiments*. 2nd ed. New York: Springer.

SAS. 2021. *JMP Pro 15*. https://www.jmp.com/en_us/home.html, accessed May 20th, 2022.

Wang, H., Q. Xiao, and A. 2020. "Musings About Constructions of Efficient Latin Hypercube Designs with Flexible Run-sizes". *arXiv preprint arXiv:2010.09154*.

Zhang, A., H. Li, S. Quan, and Z. Yang. 2018. *UniDOE: Uniform Design of Experiments (CRAN)*. https://CRAN.R-project.org/package=UniDOE, accessed May 21st, 2021.

AUTHOR BIOGRAPHIES

THOMAS W. LUCAS is a Professor of Operations Research at the Naval Postgraduate School (NPS) in Monterey, California. His primary research interests are simulation, design and analysis of computer experiments, warfare modeling and analysis, and robust Bayesian statistics. His email address is twlucas@nps.edu and his webpage is http://faculty.nps.edu/twlucas/.

JEFFREY D. PARKER is a Lieutenant Colonel in the United States Marine Corps. He received his B.S. in Quantitative Economics from the United States Naval Academy. He earned an M.S. and Ph.D. in Operations Research from the Naval Postgraduate School. His primary research interests are simulation, optimization, and design and analysis of large-scale computer experiments. He works at Headquarters Marine Corps, Combat Development and Integration evaluating innovative technologies, integrating processes, and translating leadership vision into capabilities to modernize the Marine Corps and enable the Joint/Combined Force. His email address is jeffrey.d.parker@usmc.mil.