Received: 26 January 2018 | Revised: 14 August 2018 | Accepted: 18 September 2018

DOI: 10.1002/qre.2413

SPECIAL ISSUE ARTICLE

WILEY

Selecting a near-optimal design for multiple criteria with improved robustness to different user priorities

Sarah E. Burke[1] | Lu Lu[2] | Christine M. Anderson-Cook[3] | Douglas C. Montgomery[4]

[1]Scientific Test and Analysis Techniques Center of Excellence, The Perduco Group, Dayton, OH, USA

[2]Department of Mathematics and Statistics, University of South Florida, Tampa, Florida, USA

[3]Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, New Mexico, USA

[4]School of Computing, Informatics and Decisions Systems, Engineering Department of Industrial Engineering, Arizona State University, Tempe, Arizona, USA

Correspondence
Sarah E. Burke, The Perduco Group, Wright-Patterson AFB, OH 45433, USA.
Email: seburke89@gmail.com

Funding information
Science of Test Research Consortium, Grant/Award Number: PSA-ASU-ODEX-16-04

Abstract

In a decision-making process, relying on only one objective can often lead to oversimplified decisions that ignore important considerations. Incorporating multiple, and likely competing, objectives is critical for balancing trade-offs on different aspects of performance. When multiple objectives are considered, it is often hard to make a precise decision on how to weight the different objectives when combining their performance for ranking and selecting designs. We show that there are situations when selecting a design with near-optimality for a broad range of weight combinations of the criteria is a better test selection strategy compared with choosing a design that is strictly optimal under very restricted conditions. We propose a new design selection strategy that identifies several top-ranked solutions across broad weight combinations using layered Pareto fronts and then selects the final design that offers the best robustness to different user priorities. This method involves identifying multiple leading solutions based on the primary objectives and comparing the alternatives using secondary objectives to make the final decision. We focus on the selection of screening designs because they are widely used both in industrial research, development, and operational testing. The method is illustrated with an example of selecting a single design from a catalog of designs of a fixed size. However, the method can be adapted to more general designed experiment selection problems that involve searching through a large design space.

KEYWORDS

DMRCS, *D*-optimality, factor correlations, G-optimality, I-optimality, Pareto front, power, projections of designs

1 | INTRODUCTION

When selecting a designed experiment to screen the impact of potential factors on one or more responses of interest, the objectives of the test or study need to be carefully considered and matched with criteria used to evaluate the appropriateness of the design. Screening experiments are generally used early in a testing regime that employs a sequential experimentation strategy to explore a moderate to large number of factors. The goal is to determine which factors have the largest impacts on the response(s), with first-order effects and two-factor interaction effects being of primary interest. Quadratic effects, which capture curvature in the response surface, may also be of some interest. In operational testing environments, which often involve many factors including design factors, environmental and nuisance factors, and factors that are difficult to change, cost-effective and tailored designed experiments are a necessity. Resources are often at a premium in these experiments because of the high-cost nature of these tests. It is important to have a design selection strategy that maximizes the information learned in the experiment while satisfying several different competing priorities given the experimental resources, which include the number of runs, time, equipment, and operators available.

In this paper, we illustrate a multiple criteria assessment strategy for evaluating and prioritizing designs from a design catalog. This strategy will ensure that the best design (that which matches the specific priorities of the test) can be selected. We first describe several useful categories of objectives to evaluate designed experiments. We then define quantitative criteria to assess what constitutes a best design based on how well each design fulfills the requirements for that objective. When considering multiple criteria, the best choice could change depending on how we prioritize or weight each of the criteria. A common strategy to balance different priorities is through a desirability or utility function.[1] Desirability functions (DFs) combine quantitative metrics of the different objectives into a single expression using weights to reflect the priority of each metric. A user-specified weight combination on the criteria specifies how each criterion is prioritized relative to the others. Therefore, we discuss the trade-offs between an optimal, local optimal, and near-optimal design. We define an optimal design as the design that is optimal for a single criterion. A local optimal design is one that is optimal for a particular weight combination of the multiple criteria combined into a DF. A near-optimal design is one that, while not necessarily strictly optimal, is among the leading (top-ranked) choices across a broader range of weight combinations (as opposed to a particular combination) for the chosen criteria. The optimal and local optimal designs have been extensively studied in the literature. This paper proposes a new near-optimal design strategy to explicitly evaluate the impacts of weighting choices on the design selection and the trade-off between design superiority within more narrowly restricted weight regions with a focus on designs with near-optimality across broader user priorities. Different people using the results from experiments often have divergent views of the importance of the different objectives of the test. For example, leadership in the Department of Defense, a program manager, and a program's chief tester may not all agree on the priorities in an operational test. We show that there are situations when selecting a design with near-optimality for a broad range of weight combinations of the various criteria is a better test selection strategy compared with choosing a design that is strictly optimal under very restricted conditions. This strategy offers more balance across divergent views rather than emphasizing the optimality for a narrow view. Overall, in many situations it can be beneficial to not focus too narrowly on either a single criterion or on a specific weight combination of a small set of criteria, as these approaches may not lead to a design robust to unanticipated results when the experiment is actually implemented.

Pareto fronts (PFs)[2] have been used to identify a top design based on two, three, or four criteria for a variety of design scenarios.[3-5] Ideal candidates were specified as those that were best for a range of weights representative of general test priorities. This strategy works well for experiments when consensus can be reached about common goals but is difficult when divergent objectives cannot be consolidated to an agreed-upon region of weights for the objectives. In these cases, looking more broadly at near-optimality, here defined as those choices that rank in the Top N positions (as opposed to the best, ie, highest-ranked choice) for a broad range of weight combinations, should be considered. Note we focus on the design selection of a fixed design size, ie, the total number of runs. We use N to denote the number of top designs to evaluate using the proposed method to choose the final design. The Top N strategy for multiple objective decision-making was first introduced by Burke et al[6] to select multiple leading solutions by considering their robustness to different weight combinations. The method identifies the Top N layers of PFs, which form multiple groups of ranked solutions with the top layer containing solutions strictly better than the second layer on all objectives and so on. Burke et al[6] demonstrated the Top N layers of PFs must include the Top N solutions for any possible weight combinations and hence offer an objective set of solutions before evaluating the weight impact and robustness. Burke et al[6] illustrate how the Top N strategy (described in Section 2.2) can be effectively used to identify multiple top contenders when the goal is to choose several results, such as the most critical stockpiles to receive additional funding. This approach can also work well in the context of designed experiments even though only a single experiment will ultimately be run. In this case, looking at the top choices across a variety of prioritizations of the design criteria can give insights into the robustness of multiple-purpose designs that are able to perform well across all the objectives of the test. Graphical methods are used to visualize the alternative design choices and facilitate discussions between stakeholders so they can reach consensus about which design best balances the priorities. Since stakeholders involved in the decision-making process may have diverse quantitative and statistical skills, graphical methods are accessible and allow equal participation during discussions. In addition, since some characteristics of good designs are difficult to compress into single-number summaries, the strategy of highlighting a small set of leading candidates to examine more closely with qualitative or high-dimensional summaries is advantageous. This matches the strategy described by the Define-Measure-Reduce-Combine-Select (DMRCS) process[7] and positions the experimenter to better understand leading alternatives, while keeping the decision-making process manageable.

The method described is general in that the criteria considered, how they are combined into a quantitative summary through a DF, and how close to optimal the experimenter wishes to stay can all be chosen separately and are based on the given test. This paper illustrates the method for the catalog of designs introduced in Schoen et al[8] and further studied in Jones et al[9] for a two-level five-factor 24-run screening experiment, considering cases with a wide variety of objectives. The categories of criteria considered are (a) good estimation of model parameters for models of various complexity, (b) good prediction capabilities throughout the design space, (c) power to achieve statistical significance when testing factor effects, (d) correlation between terms in the model, (e) bias for estimation of terms in the model and/or natural variability, and (f) performance of the designs if some factors are not active and projections to lower dimensional design spaces are considered. Numerical summaries of the different criteria for each of these categories are provided for all designs in the catalog. We explore designs that rank in the top three across ranges of weights when simultaneously considering three design criteria. The method is general enough to be applied to seeking any Top *N* designs based on considering any set of chosen criteria. However, the authors recommend a careful examination when choosing the set of criteria because considering too many criteria simultaneously will lead to too many possible choices to consider and often mediocre solutions. The particular choices of which criteria and how narrowly to consider near-optimality can easily be adjusted based on the priorities of the experimenters.

The top three designs for an assortment of sets of criteria are given, and discussion is provided for several cases where the choice between "best (top one choice) for a smaller set of criteria weights" is contrasted with "nearly-optimal (among the Top *N* choices) for a larger set of criteria weights." We hope that providing both the raw summaries for the collection of designs as well as top choices for some common groups of objectives will allow the experimenter to choose a design that most appropriately matches their priorities in a particular study. Computing tools for replicating the investigations summarized in this paper for other scenarios and collections of designs are also provided. Section 2 provides details about the criteria that were considered as well as some background on PFs, layered PFs, and the DMRCS decision-making process, which are the key elements in the proposed design selection process. Section 3 illustrates the method for the catalog of designs for five-factor 24-run screening experiments. Section 4 describes the JMP and R tools available for implementing the described methodology for the general design selection when considering multiple criteria simultaneously. Section 5 provides discussion and conclusions for applying the methodology to other scenarios.

2 | BACKGROUND

In this section, we provide details of the considered set of criteria to evaluate screening designs, as well as some background on formal approaches for quantitatively examining multiple criteria across a variety of different priorities.

2.1 | Criteria

For general design evaluation and selection, we recommend exploring a large number of design criteria that cover different aspects of design performance as well as different possibilities of model complexity, including terms for main effects, two-factor interactions, quadratics, and even third-order terms. Note for our particular example involving only screening experiments with two levels for each design factor, there is no capability for estimating quadratic or higher order terms. Since the general methodology and the developed computational tools are suitable for broader applications than the screening experiments discussed in this paper, we keep our discussion about the method and the tools at a general level for now, and then focus on a particular subset of criteria for the case study on screening experiments. Even though the collection described is not an exhaustive list of all possible design criteria that might be of interest to all experimenters, it does offer a broad spectrum of common criteria applicable to many screening design applications, particularly in the science of test and developmental or operational testing environments.

We use the catalog of 63 non-isomorphic five-factor 24-run designs from Schoen et al[8] as an illustrative example. We organize our criteria into six categories based on the different emphases of the evaluated criteria on various aspects of design performance, which include model parameter estimation precision, prediction variance, power, correlation, bias, and projection. To facilitate discussion of the details of the metrics, we first introduce basic notation for the definitions. Since we explore design performance for models with different numbers of terms and complexity, we first define the model matrices consisting of different groups of model terms.

We denote model matrices consisting of only the main effects ($\boldsymbol{X}_1$) and two-factor interactions ($\boldsymbol{X}_2$). Note $\boldsymbol{X}_1$ also includes the intercept. Let p_1 and p_2 denote the number of terms included in the model matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, respectively. For the designs with five factors, we have $p_1 = 6$ and $p_2 = 10$. We also define $\boldsymbol{X}_{12} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ as the model matrix for the first-order plus two-factor interactions model. We use $p_0 = p_1 + p_2$ to denote the total number of parameters in $\boldsymbol{X}_{12}$. The design size is denoted by n, which equals 24 for all of the designs in the catalog for the illustrative example. On the basis of a comparison of the sample size $n = 24$ with $p_1 = 6$ and $p_2 = 10$, we can see that it should theoretically be possible for the screening designs in our catalog to estimate all of the main effects and two-factor interactions. However, since the designs are all at 2-levels, it is not possible to estimate quadratic or higher order terms. Considering this larger design size compared with the 2_{V}^{5-1} 16-run design offers extra degrees of freedom for evaluating the error variance.

The first category of design criteria focuses on the precision of estimated model parameters. This objective is typically considered one of the primary considerations when choosing a screening design. It is important to be able to identify which terms in the model are most influential on the response and to have minimal uncertainty associated with the estimated effects. If this is part of sequential experimentation, then choices are often made about which factors to explore further and which to remove. We consider D-optimality and A-optimality criteria[10 (p 469)] for the main effects only model (denoted by M_1) and for the first-order plus two-factor interactions model (M_2). The D-criterion measures the overall precision of all model parameters based on the total volume of the confidence region for all model terms, while the A-criterion only measures the sum of the variances of individual coefficient estimates. Instead of using the original scales for these metrics, we calculate the D- and A-efficiency, which measure the relative performance of a particular design relative to the best performance on the chosen criterion for designs of the same size. We use D_1 and D_2 to denote the D-efficiencies for models M_1 and M_2, respectively, and A_1 and A_2 for the A-efficiencies for the same models. The criteria values are calculated by:

$$D_1 = \left|\mathbf{X}_1'\mathbf{X}_1\right|^{1/p_1} \text{ and } D_2 = \left|\mathbf{X}_{12}'\mathbf{X}_{12}\right|^{1/p_0}$$

$$A_1 = p_1/\mathrm{tr}\left\{\left(\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{n}\right)^{-1}\right\} \text{and } A_2 = p_0/\mathrm{tr}\left\{\left(\frac{\boldsymbol{X}_{12}'\boldsymbol{X}_{12}}{n}\right)^{-1}\right\}.$$

Larger D- and A-efficiencies indicate good precision for estimation.

The second aspect of design performance concerns the prediction variance. While typically a secondary characteristic in screening experiments, this aspect of the design can be important if the model will be used to predict the outcomes of new observations. We consider the I- and G-optimality criteria[10 (p 471–473)] for both M_1 and M_2. Both criteria are based on evaluating the scaled prediction variance (SPV). For example, SPV $= n\boldsymbol{x}_1'\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{x}_1$ for model M_1, where $\boldsymbol{x}_1$ is the location in the input space specified in model form. The G-criterion focuses on the worst-case scenario by considering the largest SPV throughout the design space, while the I-criterion measures the average performance. Since it is generally not known a priori where new observations might be needed or where optimal operating conditions may exist, the G-criterion can be thought of as seeking to keep the worst-case scenario manageable. The I-criterion uses a more typical summary for characterizing distributions by looking at the center of the distribution.

For the G-criterion, we calculate the relative performance measured by the G-efficiency for models M_1 and M_2 which are given by:

$$G_1 = \frac{p_1}{\mathrm{Max}_{\boldsymbol{x}_1 \in R}\left\{n\boldsymbol{x}_1'\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{x}_1\right\}} \text{ and}$$

$$G_2 = \frac{p_0}{\mathrm{Max}_{\boldsymbol{x}_{12} \in R}\left\{n\boldsymbol{x}_{12}'\left(\boldsymbol{X}_{12}'\boldsymbol{X}_{12}\right)^{-1}\boldsymbol{x}_{12}\right\}}.$$

The I-criterion values for M_1 and M_2 are calculated respectively as:

$$I_1 = \frac{\int_R n\boldsymbol{x}_1'\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{x}_1 \mathrm{d}\boldsymbol{x}_1}{\int_R \mathrm{d}\boldsymbol{x}_1} \text{ and}$$

$$I_2 = \frac{\int_R n\boldsymbol{x}_{12}'\left(\boldsymbol{X}_{12}'\boldsymbol{X}_{12}\right)^{-1}\boldsymbol{x}_{12} \mathrm{d}\boldsymbol{x}_{12}}{\int_R \mathrm{d}\boldsymbol{x}_{12}},$$

where the design region of interest R is assumed to be a hypercube such that $R = [-1, 1]^5$. Note that computing the I-criterion and G-efficiency is not easily obtained in closed form for most designs and hence requires a large sample of design points across the design space to evaluate their SPVs. We used a combination of a fixed set of gridded points and a stratified random sample to cover both the corners and edges as well as the interior of the design space. Larger G-efficiency and smaller I-criterion values indicate good prediction variance. In addition to the single-number numerical criteria summarized over the design space, we also consider more informative graphical summaries, such as the fraction of design space (FDS) plot[11]. FDS plots display the entire distribution of SPV throughout the design space. This plot allows the user to see any quantile of SPV, how SPV is distributed throughout the design region, and the values of average or worst-case performance used in the optimality summaries. The *G*-efficiency and *I*-value are useful to streamline the design selection process from a large set of candidate designs while the graphical summary can be used for a more detailed comparison between the most competitive designs to choose the final winner for a particular scenario.

The third category of design criteria considers the average power for detecting different groups of design effects. This criterion summarizes the average power over each of the main effects, two-factor interaction terms, and over the collection of both main and interaction effects. Power measures the probability of detecting a design effect given that there is an effect. Power is dependent on the size of the effect with generally higher power for detecting a larger effect. In many design situations, the experimenters may have intuition about what size of effects are thought to be important. Tailoring the power summaries to match these expectations can help with a priori understanding of how effective the design will be. We explore different scenarios with different sizes of design effects measured by the signal-to-noise ratio (SNR), the ratio of the effect size to detect (δ) and the experimental error (σ).

The power for the jth design effect of size δ with an SNR $r = \frac{\delta}{\sigma}$ is calculated as

$$\text{Power}_j^r = P\left\{F\left(1, n - p_0, \text{ncp} = \frac{\delta^2}{2\sigma^2\left(\boldsymbol{X}_{12}'\boldsymbol{X}_{12}\right)_{j,j}^{-1}}\right) > F_{1,n-p_0,\alpha}\right\}.$$

We use $\text{Power}_M^r = \sum_2^{p_1}\text{Power}_j^r/(p_1 - 1)$ to denote the average power over the main effects at an SNR r. Similarly, the average power for the two-factor interactions and all of the main plus interaction effects are denoted by $\text{Power}_{\text{T}}^r$ and $\text{Power}_{\text{MT}}^r$, respectively. Since it is generally not known a priori which effects in a screening experiment are active, the average power allows for a quick summary of what is expected from the design across a given category of terms. In the summary of results for the example in Section 3, we explore different effect sizes with $r \in \{1, 2, 3\}$. Larger average power indicates a better chance to correctly identify important design effects for the interested group.

Another important aspect considers the aliasing structure of the design matrix, which is measured by the absolute correlation matrix:

$$|\text{Cor}| = |\text{D}_{\text{cov}}\text{Cov}\text{D}_{\text{cov}}|,$$

where $\text{Cov} = \boldsymbol{X}'\boldsymbol{X}$ is the covariance matrix of the model coefficients, $\text{D}_{\text{cov}} = \text{diag}\left\{1/\sqrt{\text{Cov}_{j,j}}\right\}$, and $|\cdot|$ takes the absolute value of each entry of a matrix. In general, it is possible to explore the average correlation within each group of effects including main effects, two-factor interactions, pure quadratics, and third-order effects for a general model involving relevant terms. Only the main effects and two-factor interactions are relevant for the two-level screening designs in our example. We look at the average correlation between different groups (here main effects and two-factor interactions) and also the average correlation across different groups. Smaller average correlations are generally preferred as this indicates that the relevant design effects are less confounded with other model terms, and the postexperiment analysis is more likely to be able to separate out the estimates of individual effects. The average correlations give convenient compact quantitative summaries over groups of effects. However, more information can be shown in the color maps of correlation, which display the absolute correlation between all pairs of individual model terms. In our design selection approach, we use the numerical summaries (the average correlations) to streamline the selection process for a large number of designs and then use the more descriptive correlation maps to further examine the leading candidates to guide selection of the final design.

Considering that in many experiments we are unsure about the exact form of our specified model, we may want some protection against potential bias of the model estimates in case the model we chose to fit was incorrect. The following criteria quantify the potential biases on the estimated model coefficients as well as experimental error variance if the specified model is in fact missing some important terms. We first consider this in the context of examining the impact of active two-factor interactions on the estimates of the main effects. The criterion $\text{tr}(\boldsymbol{A}\boldsymbol{A}')$, alias matrix $\boldsymbol{A} = \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2$, measures the bias on the estimated main effects when at least one two-factor interaction is active[10 (p 373)]. The bias on the estimated experimental error variance[10 (p 374)] in this case is quantified by $\text{tr}(\boldsymbol{R}'\boldsymbol{R})$, where $\boldsymbol{R} = \boldsymbol{X}_1\boldsymbol{A} - \boldsymbol{X}_2$. Since the values of the coefficients for omitted terms in the model are almost always unknown, smaller values indicate smaller potential bias on the quantities of interest by a certain form of model misspecification.

The last important aspect to consider for design selection is the projection properties to lower numbers of factors. For screening experiments, it is common that only a subset of the design factors initially explored are actually active. In addition to wanting good design performance for the full dimension space (eg, all five factors in our example), we also want good performance when it is projected down to a lower dimension with only a subset of active factors. When there is large uncertainty in the number of active factors, it is recommended to seek a balanced performance across multiple possibilities rather than optimizing based on a single guessed scenario. In our example, we consider the design performance projected down to all subsets of four or three design factors. To obtain the summary of performance in lower dimensions, we average over results from all potential projected models (five models for four factors and 10 models for three factors depending on which factors are assumed to be active) for all criteria chosen from the above-mentioned criteria (23 for the full five-factor case) across all categories of potential objectives. In the end, each design in the catalog is evaluated based on a total of 69 criteria (23 in the full dimension and 46 for projections) on various aspects of design performance.

2.2 | Pareto fronts, Top N choices, and ties

Since there are multiple competing objectives for many experiments, it is important to have a strategy for objectively removing noncontenders from consideration while not overlooking a promising option. One approach to multiobjective optimization is Pareto optimization: a technique used to optimize multiple criteria simultaneously while incorporating the experimenter's priorities for the study. A PF identifies a collection of competitive solutions that can be examined before making a final choice. The PF is formed by nondominated solutions,[12] those solutions that are not outperformed by any other solution based on all the criteria. In other words, for a solution on the PF, there is no other solution that is at least as good on all criteria and strictly better on at least one criterion. Since the selection of the PF does not rely on the specification of weights, scaling, or DFs for combining the criteria, it offers an objective set of superior solutions for further exploration. The method is divided into two phases of analysis: (a) an objective phase where potential rational solutions are identified and noncontending choices, which will never emerge as ideal, are eliminated and (b) a subjective stage where the remaining solutions are evaluated for different prioritizations of the criteria. A final decision is made by evaluating the results from phase two of the analysis through the use of several graphical tools[2-4,13]. Anderson-Cook and Lu[7] propose a five-step process DMRCS for decision-making with competing objectives. PF optimization is an ideal method used in the Reduce step of this process.

Figure 1 shows a representation of a PF for a scenario with two criteria, both optimized by minimizing. Ideally, we would like a solution at the utopia point, the point that has the best possible values for all criteria simultaneously (represented by the triangle in Figure 1). However, this ideal solution rarely exists in practice since there are generally trade-offs in the criteria. The solutions connected by the line in Figure 1 make up the PF: the set of nondominated points in the criterion region, with the remaining points dominated by at least one PF solution.

Constructing the PF is straightforward from an enumerated list of solutions, with several software packages having an implementation of this automatically. If the PF needs to be constructed within a search algorithm of candidate solutions, the Pareto aggregating point exchange (PAPE) algorithm[2,12] and its enhancements[13,14] for design selection have been used.

The first phase of the PF optimization algorithm eliminates noncontending solutions, simplifying the problem and making the final decision more manageable. This phase is independent of any differences in measurement scales of the criteria because the algorithm simply identifies the dominant solutions based on their evaluated criteria values. An advantage of PF optimization is that it separates the objective and subjective phases and allows decision-makers to bring their own priorities only into the subjective phase. Decision-makers then examine the optimal solutions for different prioritizations (or weightings) of the criteria only after the initial set of solutions have been objectively reduced. Because of this flexibility, the method can accommodate several competing sets of priorities, particularly when multiple decision-

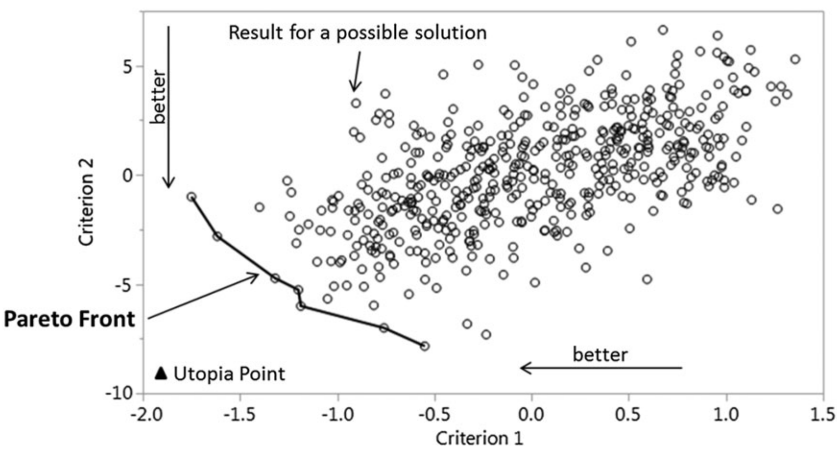

FIGURE 1 Graphical representation of a Pareto front for two criteria

makers are involved. The subjective phase also lets the user evaluate the sensitivity of a particular solution based on different prioritizations on the criteria.

The algorithm can be adapted to accommodate any number of criteria. However, we note that the computation time increases exponentially as the number of criteria increases. It is therefore recommended that practitioners be selective on the number of criteria used for optimization. In our example, we restrict our exploration of leading designs to simultaneously consider only three criteria at a time. Including too many criteria not only reduces the computational efficiency but also can result in mediocre solutions due to large trade-offs associated when considering too many criteria.

Pareto optimization helps identify an optimal solution based on user priorities and balancing multiple objectives. In this paper, the goal is not to identify just an optimal solution but also to consider near-optimal solutions that are robust compared with the other designs across a wider range of weight combinations of the criteria. A design that is optimal for one criterion could have poor performance for another criterion. For example, a *D*-optimal design offers the most precise estimation of the specified model but may provide little protection if the specified model is far from the true model. Another design may offer compromised performance on both criteria but better robustness when a decision needs to accommodate different opinions on how much each criterion should be valued in the design selection process. To accommodate the desire to consider near-optimal solutions in the design selection, we have adapted the PF approach[2] to include the Top *N* solutions, not just those which are strictly best. Layered PFs[6] allow all designs that could rank in the Top *N* choices to be identified as they consider potential solutions that lie just behind the PF as potential candidates. Considering optimal and near-optimal designs provides a larger pool of potential solutions to evaluate. Designs under consideration may not be strictly optimal; however, they may have good performance robust to the choice of the weight combinations on the criteria considered. This is particularly beneficial when multiple decision-makers, with differing opinions, are part of the design selection process. A larger pool of potential solutions also allows subsequent exploration of qualitative or higher dimensional summaries (such as the FDS and correlation plots) to be included in discussions. The inclusion of all Top N layered PFs ensures that any potential near-optimal solution will be considered and that their ranking for any combination of weightings are included when robustness is evaluated.

The Top N Pareto Front Search (TopN-PFS) algorithm has been programmed in the JMP Scripting Language (V. 13) and is available as an add-in in JMP[6]. This functionality is also built in an R package named multiple criteria design selection (MCDS) for implementing the methodology proposed in this paper and is discussed more in Section 4. This algorithm can be used to identify the Top *N* solutions in an enumerated list. The PF must necessarily contain all solutions that are best for any weight combination of an L_p-norm form of a DF[2]. By extension, a Top *N* solution for these forms of DFs will be in the Top *N* layers of the PF. For example, the second best solution for a given weight choice of a DF must lie on either the first or second PF layer. Hence, reducing the candidate solution set to just consider the Top *N* layers of the PF will provide an objective set of superior solutions that must contain the Top *N* solutions without relying on any subjective choices on user priorities.

In the first phase of the TopN-PFS in the JMP add-in, the PF layers are built by finding groups of nondominated points. In particular, the points on the traditional PF are labeled as belonging to layer 1 and then set aside. The next PF layer is built by examining the remaining points. The nondominated points on the PF for this reduced set are labeled as belonging to layer 2 and again set aside. This process continues until the Top *N* layers have been identified. The R MCDS package uses a method employed in the nondominated sorting genetic algorithm (NSGA-II)[15-17] for ranking a population of solutions based on layers of PFs. Although the two algorithms use slightly different mechanisms, they serve the same purposes for finding the Top *N* layers of PFs from which to select the Top *N* solutions. This phase of the algorithm corresponds to the Reduce step in the DMRCS process[7].

In the second subjective stage, DFs[1] are used to rank the points from phase 1 for different combinations of weights on the criteria. The different weights allow users to compare changes in the solutions when changing the prioritization of the different criteria. However, the criteria, which measure different aspects of the potential choices, are likely measured on different scales. To allow for a fairer comparison of each criterion, values x_{ij}, denoting the j^{th} value in the dataset for criterion i, are scaled to values z_{ij} such that

$$z_{ij} = \begin{cases} \dfrac{x_{ij} - \text{worst}_i}{\text{best}_i - \text{worst}_i}, & \text{if criterion } i \text{ is maximized,} \\ \dfrac{\text{worst}_i - x_{ij}}{\text{worst}_i - \text{best}_i}, & \text{if criterion } i \text{ is minimized,} \end{cases}$$

where worst_i and best_i are defined as the worst and best values for the i^{th} criterion, respectively. These scaled values are called "desirability scores." Using these scaling schemes, a value of 0 corresponds to the worst value and a value of 1 corresponds to the best value, regardless of the original scaling or if the goal was to minimize, maximize, or hit a target (which is turned into a problem for minimizing the distance to the target). How $best_i$ and $worst_i$ are defined depends on the judgment of the decision-maker or on the properties of the particular data set under consideration. In the JMP add-in that was developed in Burke et al[6], four options are available:

(1) The best and worst values are calculated from the data only in the layered PFs.
(2) The best and worst values are calculated from the entire dataset.
(3) The best value is calculated from the data in the layered PFs and the user provides the worst values for each criterion.
(4) The user provides the best and worst values for each criterion.

In the case of scaling options 1 or 2, for example, if criterion i is maximized, the worst value would be the minimum in the applicable dataset, while the best value would be the maximum. If criterion i is minimized in these scaling scenarios, the reverse is true. In Section 3, we focus on the first two options for identifying the top choices of designs, which are also the choices available in the R package. However, since the original values for all of the criteria are included in tables, implementing alternate scaling is straightforward in the JMP add-in.

The third or fourth options for scaling may be appropriate when user-specified reference values are available or if the ranges of the criteria are vastly different. For example, if the range of criterion 1 covers excellent values to very poor values while the range for criterion 2 covers good to fair values, the user may want to supply values for $best_i$ and $worst_i$ to balance these ranges. The choice of scaling can have a substantial impact on the final designs recommended from the TopN-PFS approach. Hence, careful thought should be given to ensure that the choice of scaling matches the decision-maker's intentions. A sensitivity study should be conducted to understand the impact of the different scaling options.

With the original data transformed to desirability scores z_{ij} (values between 0 and 1), we can fairly combine these values and compare them across different weighting schemes on the criteria. Two common DFs (the additive and multiplicative DFs) are considered for k criteria:

$$\text{Add DF}_j = \sum_{i=1}^{k} w_i z_{ij},$$

$$\text{Multi DF}_j = \prod_{i=1}^{k} z_{ij}^{w_i},$$

where the weights $w_i \geq 0$ satisfy $\sum_{i=1}^{k} w_i = 1$, and z_{ij} is the desirability score of the jth solution for the ith criterion. The choice of DF depends on the priorities of the user. The multiplicative DF penalizes low criterion values more than the additive DF. For example, if one choice performs very poorly for one of the criteria, it is very difficult for other criterion scores to compensate for this performance. The additive DF, on the other hand, is more forgiving of low criterion scores since a very high value for one criterion can override a low value for another criterion. The choice of DF can impact which solutions are highlighted as best with this algorithm. The effect of the form of the DF was considered in a simulation study[18].

DF scores are calculated for each solution identified in phase 1 of the algorithm and across different combinations of weights on the criteria that match the user's priorities. For each weight combination, the DF scores are ranked from highest to lowest to identify the Top N solutions. This stage of the algorithm corresponds to the Combine step in the DMRCS process. Graphical summaries for comparing and selecting solutions are illustrated for the examples shown in Section 3.

One interesting consideration, particularly in the design of experiments (DOE) application, is how to handle ties. Without careful handling of ties, it would be easy to lose track of potentially desirable solutions. Two types of ties can occur in the TopN-PFS algorithm. The first type occurs in the first phase and occurs when two or more solutions have exactly the same values for all criteria. No matter how we scale the data, these solutions always have the same values of the criteria. Since we wish to include any potential solution that performs well for the criteria, we include all tied solutions. Therefore, to handle ties we choose a representative solution while building each layer in the PF. After that layer of the PF has been constructed, all tied solutions in that layer are added to that layer.

The second type of tie occurs in the second phase of the TopN-PFS algorithm when we calculate DF scores for all solutions. More than one solution may have the same DF score for a particular weight combination of the criteria even though all the criteria values are not identical. Therefore, new ties might be introduced at this stage for a given weight combination. If two or more solutions have the same DF score and are in the Top N, these tied solutions are given the same ranking. More than N solutions could be identified for any weight combination from either source of ties. Once at least N solutions have been identified, no additional solutions are included in the ranking.

3 | FIVE-FACTOR 24-RUN EXAMPLE

To illustrate the method for selecting near-optimal solutions, we use one of the catalogs of two-level screening designs for five factors in 24 runs, enumerated in Schoen et al[8] and further described in Jones et al.[9] The catalog has 63 nonisomorphic designs, and we refer to the designs with the numbering from the original catalog. These designs are strength-2 orthogonal arrays (OAs) that can be used as alternatives to traditional screening designs, such as a full factorial or fractional factorial design. Strength-2 OAs are designs such that any combination of two columns in the design matrix have all possible combinations of factor levels occur an equal number of times. For example, for factors $X1$ and $X2$, the factor level combinations (1,1), (1, −1), (−1, 1), and (−1, −1) occur the same number of times in the design. Strength-2 OAs can fit a main effects model (M_1) with minimum variance[19], but unlike factorial designs, they may not always be able to estimate a main effects and two-factor interactions model. If the OA can fit this model, it may not be with minimum variance, and the model terms may be aliased. As discussed previously, model M_2 for five factors has 16 terms. An advantage of using the 24-run OA for five factors compared with the 2_{V}^{5-1} fractional factorial design is that the OA has sufficient degrees of freedom to evaluate all model terms as well as the error variance. Jones et al[9] investigated the properties of strength-2 OAs for four to six factors. They restricted their criteria for evaluation to D- and A-efficiency.

To begin the exploration of the designs in the catalog, R code was used to calculate summaries across the categories described in Section 2. Of the 63 designs, only 36 were able to estimate the full main effects with two-factor interactions model (M_2), and hence we restrict our reporting to those designs, listed in Table 1 with each design numbered as it was in Schoen et al.[8] Note that for some of the criteria, such as AC_T in Table 1, there are several designs tied for the best available value. This is not uncommon in DOE applications, and appropriate handling of ties is essential. The strategy of eliminating clear noncontenders from further consideration matches with the strategy described in the Reduce step of DMRCS[7] to eliminate choices that are not suitable solutions. The best value across all designs for each of the criteria is highlighted in bold. Note that Design 4 is best for a large number of the criteria, including highest D-, A-, and G-efficiencies and lowest I-criterion value. It is also best for several of the power summaries and the average absolute correlation among two-factor interactions. If the goal was strictly to consider the best estimation of model parameters or prediction throughout the design space for the five-factor scenario, a strong argument could be made that Design 4 is an ideal choice.

In the Appendix, we include results similar to those in Table 1 for the case when a reduced model involving only four (Table A1) or three (Table A2) factors are active. These results average across the five or 10 subsets, respectively, as one or two factors are removed. Table 2 summarizes the range of values observed across the 36 contending designs for each of five, four, and three factors active. In general, as the number of factors (and terms) in the model is reduced, the attainable values improve for all criteria. Design 4 continues to be one of the most attractive designs in terms of its performance on the D-, A-, G-efficiencies, and many of the power and correlation summaries. However, Design 1 is more attractive if reduced models with only four active factors are considered. In this case, Design 1 ties with Design 4 for almost of the criteria for which it is optimal except the average power for two-factor interactions and the average correlation between interactions. In addition, Design 1 is optimal for other criteria including the I-criterion, average power for main effects, and the average correlation between main effects and two-factor interactions. Besides Designs 1 and 4, Designs 2 and 3 are also quite competitive for several criteria. When reduced models with only three active factors are considered, Design 1 appears to dominate for almost all criteria except $\mathrm{tr}(\boldsymbol{R}'\boldsymbol{R})$. Note the final design selection should be based on considering the criteria that are most relevant to the particular experiment. For example, if subject matter expertise suggests all design factors are likely to be useful, then the decision should be made based on the design criteria of interest summarized over the full dimension of the five-factor input space.

Recall that each screening experiment has distinct objectives, and which criteria the experimenter focuses on should be matched to the testing goals. An experimenter might naively suggest that since all of the criteria described in Section 2 are potentially relevant to a good screening experiment, all of the criteria should be included when constructing a high-dimensional PF. JMP software[20] allows a PF to be constructed based on as many criteria as desired. For this example, all 36 of the contending designs are included on the PF when including all the criteria, which clearly does not help with highlighting best designs on which to focus for making the decision of which experiment to run. This illustrates the importance of choosing the relevant criteria for a particular testing scenario to ensure that the most important priorities of the experiment are well satisfied by the chosen design. Selecting a large number of criteria leads to both mediocrity in performance of some of the criteria and a very large number of contending designs remaining under consideration.

Hence for the remainder of the discussion, we restrict ourselves to scenarios where the experimenter has carefully selected three criteria on which to focus. Both the R code and the TopN-PFS JMP add-in focus on solutions

TABLE 1 Selected criteria values for the Schoen et al[8] catalog of five-factor designs with 24 runs

Design	*D*-efficiency	*A*-efficiency	*I*	*G*-efficiency	pwr_M^2	pwr_T^2	pwr_{MT}^2	AC_T	$AC_{M\times T}$	AC_{MT}	$tr(AA')^{p4}$	$tr(R'R)^{p4}$
1	0.868	0.719	0.171	0.457	**0.857**	0.664	0.729	0.111	**0**	0.048	**0.00**	240
2	0.926	0.842	0.162	0.571	0.792	0.792	0.792	**0.022**	0.04	**0.029**	0.67	224
3	0.902	0.792	0.164	0.400	0.827	0.736	0.766	0.067	0.02	0.038	0.33	232
4	**0.939**	**0.884**	**0.156**	**0.792**	0.810	**0.810**	**0.810**	**0.022**	0.04	**0.029**	0.67	224
6	0.800	0.563	0.211	0.299	0.770	0.587	0.648	0.067	0.04	0.048	0.67	224
7	0.874	0.736	0.177	0.461	0.785	0.711	0.736	0.067	0.04	0.048	0.67	224
8	0.902	0.792	0.172	0.400	0.756	0.771	0.766	**0.022**	0.06	0.038	1.00	216
9	0.766	0.486	0.241	0.169	0.704	0.528	0.587	0.067	0.06	0.057	1.00	216
10	0.842	0.653	0.197	0.294	0.729	0.664	0.686	0.067	0.06	0.057	1.00	216
11	0.812	0.595	0.215	0.287	0.675	0.631	0.646	0.067	0.08	0.067	1.33	208
12	0.738	0.443	0.271	0.242	0.600	0.492	0.528	0.111	0.08	0.086	1.33	208
13	0.813	0.604	0.213	0.337	0.682	0.642	0.655	0.067	0.08	0.067	1.33	208
15	0.646	0.215	0.529	0.080	0.341	0.280	0.301	0.111	0.1	0.095	1.67	200
16	0.902	0.792	0.173	0.400	0.751	0.774	0.766	**0.022**	0.06	0.038	1.00	216
17	0.749	0.359	0.343	0.097	0.503	0.462	0.476	0.067	0.1	0.076	1.67	200
18	0.728	0.387	0.302	0.180	0.588	0.491	0.523	0.067	0.08	0.067	1.33	208
19	0.874	0.736	0.185	0.461	0.711	0.748	0.736	**0.022**	0.08	0.048	1.33	208
21	0.812	0.595	0.217	0.287	0.667	0.635	0.646	0.067	0.08	0.067	1.33	208
22	0.812	0.595	0.229	0.287	0.595	0.671	0.646	**0.022**	0.12	0.067	2.00	192
24	0.813	0.604	0.226	0.337	0.603	0.682	0.655	**0.022**	0.12	0.067	2.00	192
25	0.749	0.359	0.343	0.097	0.488	0.470	0.476	0.067	0.1	0.076	1.67	200
26	0.842	0.653	0.209	0.294	0.642	0.707	0.686	**0.022**	0.1	0.057	1.67	200
27	0.749	0.359	0.349	0.097	0.476	0.476	0.476	0.067	0.1	0.076	1.67	200
28	0.749	0.359	0.375	0.097	0.397	0.516	0.476	**0.022**	0.14	0.076	2.33	184
31	0.733	0.412	0.307	0.216	0.496	0.496	0.496	0.067	0.12	0.086	2.00	192
34	0.868	0.719	0.189	0.457	0.703	0.741	0.729	**0.022**	0.08	0.048	1.33	208
35	0.738	0.443	0.304	0.242	0.466	0.560	0.528	**0.022**	0.16	0.086	2.67	**176**
36	0.738	0.443	0.288	0.242	0.526	0.530	0.528	0.067	0.12	0.086	2.00	192
40	0.646	0.215	0.574	0.080	0.292	0.305	0.301	0.067	0.14	0.095	2.33	184
43	0.766	0.486	0.259	0.169	0.608	0.576	0.587	0.067	0.06	0.057	1.67	200
47	0.800	0.563	0.240	0.299	0.593	0.675	0.648	**0.022**	0.08	0.048	2.00	192
48	0.728	0.387	0.324	0.180	0.508	0.530	0.523	0.067	0.08	0.067	2.00	192
49	0.728	0.387	0.349	0.180	0.473	0.548	0.523	**0.022**	0.12	0.067	2.67	**176**
50	0.800	0.563	0.243	0.299	0.580	0.682	0.648	**0.022**	0.08	0.048	2.00	192
51	0.766	0.486	0.279	0.169	0.513	0.624	0.587	**0.022**	0.1	0.057	2.33	184
52	0.766	0.486	0.282	0.169	0.496	0.632	0.587	**0.022**	0.1	0.057	2.33	184

The best criteria values are highlighted in boldface.

Abbreviations: $AC_{M\times T}$, average correlation over pairs of main effect and two-factor interaction; AC_{MT}, average correlation over pair of effects from either main effects or two-factor interactions; AC_T, average correlation over pairs of two-factor interactions; pwr_M^2, average power of main effects at $r = \frac{\delta}{\sigma} = 2$; pwr_{MT}^2, average power of main effects and two-factor interactions; pwr_T^2, average power of two-factor interactions.

with this constraint and allow two or three criteria to prioritize. With the inclusion of Tables 1, A1, and A2, it is possible to use performance on other secondary criteria when making the final decision. In addition to selecting which criteria to focus on, there are additional decisions about how to map the raw criterion values to the

TABLE 2 The table summarizes the best and worst criteria values for selected criteria

Criteria	Five-factor design		Projection to four-factor designs		Projection to three-factor designs	
	Best	Worst	Best	Worst	Best	Worst
D-efficiency	0.939	0.646	0.968	0.872	1	0.958
A-efficiency	0.884	0.215	0.936	0.737	1	0.921
I	0.156	0.574	0.240	0.311	0.346	0.375
G-efficiency	0.792	0.080	0.786	0.388	1	0.76
pwr_M^2	0.857	0.292	0.892	0.749	0.904	0.867
pwr_T^2	0.810	0.280	0.870	0.783	0.904	0.867
pwr_{MT}^2	0.810	0.301	0.870	0.770	0.904	0.867
AC_T	0.022	0.111	0.013	0.067	0	0
$AC_{M\times T}$	0	0.16	0	0.133	0	0.089
AC_{MT}	0.029	0.095	0.022	0.076	0	0.053
$\mathrm{tr}(AA')$	0	2.67	0	1.07	0	0.27
$\mathrm{tr}(R'R)$	176	240	118	144	66	72

Abbreviations: $AC_{M\times T}$, average correlation over pairs of main effect and two-factor interaction; AC_{MT}, average correlation over pair of effects from either main effects or two-factor interactions; AC_T, average correlation over pairs of two-factor interactions; pwr_M^2, average power of main effects at $r = \frac{\delta}{\sigma} = 2$; pwr_{MT}^2, average power of main effects and two-factor interactions; pwr_T^2, average power of two-factor interactions.

desirability scale and which functional form of the DF to use. Also, considering that three criteria were chosen as those of primary interest to focus on during the design selection process, each criterion should not be given too small amount of weight when evaluating their contribution to the total performance. Therefore, we constrain ourselves to weigh each criterion for at least 20% of the total weight, which leads to a maximum of 60% weight contribution for each criterion. In other words, we consider a focused weight region[21] in a central triangular region with $w_i \in [0.2, 0.6]$ in the weight space. While many choices must be made to implement the method and different weight regions may be agreed upon by different decision-makers, we find that discussions among the decision-makers lead to making informed decisions for each choice. Focusing on the weight region that covers different priorities from different decision-makers is an effective way to reach consensus and get more buy-in for the final decision.

Next we describe some scenarios that consider various groups of criteria and illustrate how different designs can emerge as potentially desirable solutions. In selecting three criteria to examine with the near-optimality approach, we select aspects of the design which balance some of the diverse goals of a good experiment[10 (p 370)]. Table 3 shows 15 scenarios of three criteria evaluated, where regardless of the method of scaling and DF employed, we end up with the same conclusion and identify the same top design. These are chosen from a large number of sets of three criteria where it was relatively common to obtain a consistent result when looking at optimality (strictly best) or near-optimality (close to best). First, we describe the details of the information summarized in Table 3. The second through fourth columns define which three criteria are being considered. The next three columns focus on optimality and identify the best three designs for maximizing the fraction of weight combinations where that design is strictly best. The final three columns focus on near-optimality by identifying which three designs have the largest fraction of weight combinations for the chosen three criteria for which the design is ranked in the top three choices.

To make this more concrete, consider the additive DF with scaling based on all the designs listed in Table 1. Case 1 prioritizes D-optimality for the five-factor (D), four-factor (D^{p4}), and three-factor (D^{p3}) scenarios. Design 4 is ranked in the top three designs for all weight combinations (100%) for $w_1D + w_2D^{p4} + w_3D^{p3}$ or $(D)^{w_1}(D^{p4})^{w_2}(D^{p3})^{w_3}$ and is strictly best for 52.4% of the weight combinations of (w_1, w_2, w_3). Tied for first place based on near-optimality (fraction of time in top three) is Design 3, but this design is never top ranked for any weight combination. This highlights the value in examining those solutions on N PF layers rather than just the first PF layer. The third best design based on near-optimality is Design 1 which is in the top three for 61.9% of the weight combinations and is strictly best for 47.6% of them. This indicates that strict design superiority is often associated with a narrower optimal weight region and hence less robustness to different user priorities. On the other hand, allowing consideration

TABLE 3 Sample scenarios when top three and top only methods select the same top design

	Optimization criteria			Optimality only			Near optimality (top three)		
Case	1	2	3	Rank = 1 (opt.%)	Rank = 2 (opt.%)	Rank = 3 (opt.%)	Rank = 1 (opt.%)	Rank = 2 (opt.%)	Rank = 3 (opt.%)
1	D	D^{p4}	D^{p3}	4 (52.4)	1 (47.6)	NA	4 (100)	3 (100)	1 (61.9)
2	A	A^{p4}	A^{p3}	4 (52.4)	1 (47.6)	NA	4 (100)	3 (100)	1 (68.4)
3	pwr_M^2	$pwr_M^{2,p4}$	$pwr_M^{2,p3}$	1 (100)	NA	NA	1 (100)	4 (100)	3 (100)
4	pwr_T^2	$pwr_T^{2,p4}$	$pwr_T^{2,p3}$	4 (96.1)	1 (3.9)	NA	4 (100)	3 (100)	2 (93.9)
5	AC_{MT}	AC_{MT}^{p4}	AC_{MT}^{p3}	2(59.3)	4 (59.3)	1 (40.6)	2 (100)	3 (100)	4 (58.9)
6	$\text{tr}(AA')$	$\text{tr}(AA')^{p4}$	$\text{tr}(AA')^{p3}$	1 (100)	NA	NA	1 (100)	3 (100)	2 (100)
7	$\text{tr}(R'R)$	$\text{tr}(R'R)^{p4}$	$\text{tr}(R'R)^{p3}$	35 (100)	NA	NA	35 (100)	49 (100)	28 (100)
8	D	I	$AC_{M \times T}$	4 (59.7)	1 (40.3)	NA	4 (100)	3 (100)	1 (57.6)
9	D	I	$\text{tr}(AA')$	4 (59.7)	1 (40.3)	NA	4 (100)	3 (100)	1 (57.6)
10	D	I	$\text{tr}(R'R)$	4 (64.9)	24 (19.1)	35 (12.6)	4 (67.5)	8 (64.1)	2 (42.4)
11	pwr_{MT}^2	AC_{MT}	$\text{tr}(AA')$	4 (87.9)	1 (12.1)	NA	4 (100)	3 (100)	2 (84)
12	pwr_{MT}^2	AC_{MT}	$\text{tr}(R'R)$	4 (67.1)	47 (26.4)	50 (26.4)	4 (76.2)	2 (66.2)	8 (55.4)
13	pwr_{MT}^2	AC_{MT}	D^{p4}	4 (100)	NA	NA	4 (100)	3 (100)	2 (100)
14	pwr_{MT}^2	AC_{MT}	D^{p3}	4 (90.5)	1 (9.5)	NA	4 (100)	3 (100)	2 (87)
15	A	G	$\text{tr}(AA')^{p3}$	4 (99.6)	1 (0.4)	NA	4 (100)	2 (97.8)	1 (90)

All the sample scenarios select the same top designs by both methods for four desirability functions (DF) and scaling combinations we have evaluated. The optimal weighting areas reported in the table were calculated using the additive DF with scaling based on all data. The superscripts $p3$ and $p4$ indicate the criteria based on the projected designs with three and four design factors, respectively.

Abbreviations: $AC_{M \times T}$, average correlation over pairs of main effect and two-factor interaction; AC_{MT}, average correlation over pair of effects from either main effects or two-factor interactions; AC_T, average correlation over pairs of two-factor interactions; pwr_M^2, average power of main effects at $r = \frac{\delta}{\sigma} = 2$; pwr_{MT}^2, average power of main effects and two-factor interactions; pwr_T^2, average power of two-factor interactions.

of near-optimality can offer more robustness and broad acceptance to the final decision. From the design selection perspective, the decision is dependent on whether the optimal weight region has comfortably covered all of the weight combinations of interest. If the 47.6% optimal weight combinations have included all sensible weight regions for the experiment, then Design 4 should be selected as the superior design. However, if a broader weight region needs to be considered for the experiment, then Design 1 can be a good choice, which offers nearly as good performance (among leading solutions) for broader weight regions of interest.

Knowing the fraction of weight combinations for which a design is optimal or near-optimal does not provide complete information. Understanding the set of prioritizations for which a design is best can help the experimenter understand if the design matches their objectives. Consider case 8 where the priorities of the experiment are to maximize D-efficiency, minimize the average SPV or I-criterion[10], and minimize the average absolute correlation between the main effects and two-factor interactions for five factors. From Table 3, we see that for the additive DF with scaling based on all of the designs listed in Table 1, Design 4 is optimal for 59.7% of the weight combinations, while Design 1 is optimal for 40.3% of weights. Design 4 is in the top three designs for near-optimality for 100% of the weight combinations, while Design 1 is only ranked in the top three for 57.6% of the weight combinations. Design 3, on the other hand is again in the top three for 100% of the weight combinations. Figure 2 introduces a new graphical summary, the design rank plot (DRP), which highlights the ranking of each design throughout all of the weight combinations, with white color for top rank, light gray for second rank, darker gray for third rank, and darkest gray for ranking below the top three. In Figure 2A, we see that Design 4 is best with a large white region when D-efficiency (bottom left) and the I-criterion (bottom right) are weighted most heavily but is ranked third when the average absolute correlation of the main effects with two-factor interactions are weighted more strongly (top). In Figure 2B, we see that the region for which Design 4 is ranked third is the set of weight combinations for which Design 1 is best. The ranking of Design 1 slips out of the top three for weight combinations where D and I criteria are valued most. Based on this plot, we can identify which design is

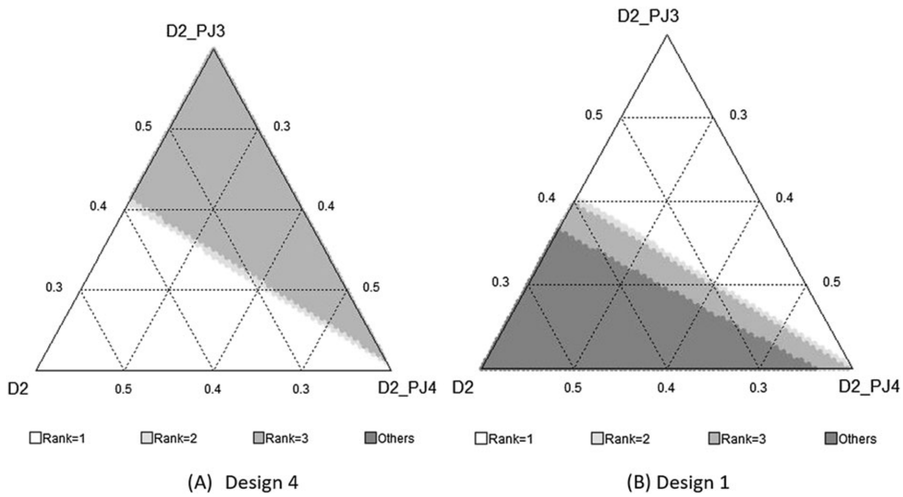

FIGURE 2 The design rank plot (DRP) for the top two designs selected for case 8 in Table 3 based on the *D*- & *I*-efficiencies for the first-order plus two-factor interaction model and the average correlation between the main and interaction effects. A, The top choice, Design 4 from the catalog, is optimal based on the additive desirability functions (DF) for 59.7% of the total weights and among the top three designs for all possible weights. B, The 2nd choice, Design 1, is optimal for 40% of the total weights but is among the top three designs for only 57.6% of the possible weights

better for the regions that are of maximal interest to the experimenter. This plot complements the synthesized efficiency plot.[3] The synthesized efficiency plot focuses on actual performance relative to the best available design for a given set of weights as opposed to rankings. Figure 3 shows the synthesized efficiency plot for case 8 where the criteria considered are *D*-efficiency, average SPV, and average absolute correlation between the main effects and two-factor interactions. In this plot, there are 20 shades of gray between white and black. Regions shaded white have efficiency that is at least 95% as good as the best available design (and must include those regions for which the design is best). The lightest shade of gray corresponds to a design that is between 90% and 95% efficient, with the next darker shade covering 85% to 90%, and so forth. Both Designs 4 and 1 are at least 85% efficient relative to the best available design across all weight combinations. The two figures provide detailed information about both the rank of the design as well as their relative performance. Overall, Design 4 is desired if precise estimation is of more interest for considering all five design factors while Design 1 is preferred if only three design factors are likely to be active.

If we examine the top designs specified in Table 3, we see that for a large number of cases, Design 4 is ranked as best. However, there are some other designs that emerge as possible choices depending on which criteria have been identified by the experimenters. In case 3, where maximum power for the main effects is of interest for the five-factor (D), four-factor (D^{p4}), and three-factor (D^{p3})

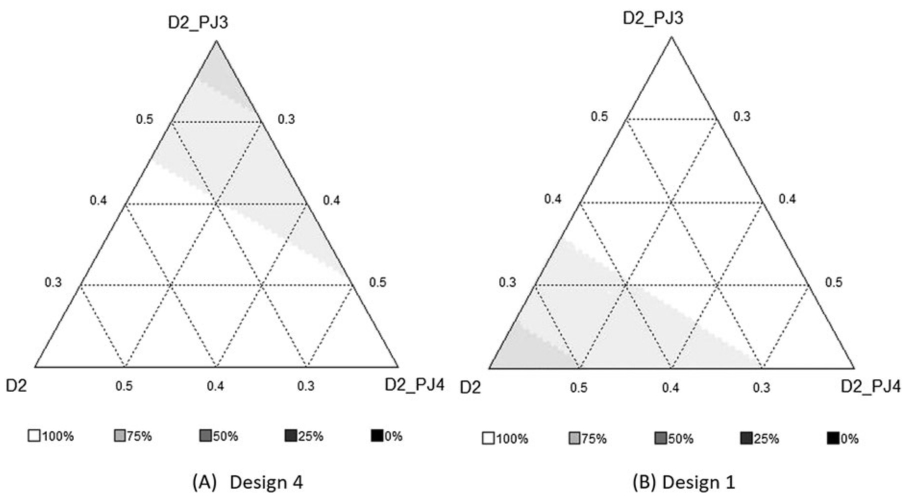

FIGURE 3 The synthesized efficiency plot (SEP) for comparing Designs 4 and 1 based on the *D*-efficiencies for the full dimension of five factors and projections down to four and three factors

scenarios, Design 1 is strictly best, with Designs 4 and 3 being near-optimal for 100% of the case but never ranked first. Case 5 identifies Design 2 as best if the priorities are to minimize the absolute correlation between the main effects and the two-way interactions. Design 35 is best if highest priority is placed on minimizing tr($\boldsymbol{R}'\boldsymbol{R}$), which would lead to minimal bias on the estimate of natural variability if the main effect model was fit and there were active two-factor interactions. Seeing how different designs perform across different options highlights that there is generally no global winner, and the final choice should be made by selecting the best choice for the experiment's priorities.

Next, we consider cases where the choice of the best design differs depending on the form of the DF and the scaling and whether we focus on near-optimality or strict optimality. This is in contrast to the results shown in Table 3 where the same designs are identified no matter which scaling method or DF were used. Some interesting cases are summarized in Table 4. The structure of the table differs slightly from that in Table 3 with the addition of two columns that describe what form of the DF and what scaling method was used. For entries in this table, more difficult trade-offs between designs may be needed to determine which design is best for a particular scenario. Consider case 1 from Table 4 where the focus is on maximizing power for all terms in the main effects with two-factor interaction model for five-, four-, and three-factor models. If we focus on optimality only, the design which is ranked first for the largest range of weight combinations is Design 1 with 63.2% of the region. Design 4 is second with 36.8% of the region. However, if we consider near-optimality, then Design 4 is ranked in the top three for 100% of the weight combinations, compared with 78.4% of the region for Design 1. In addition, the experimenter should examine for which weight combinations the designs are ranked best, second, or third. Figure 4 shows the regions where each design ranks in various places. Design 1 is best primarily when more emphasis is placed on scenarios where only three or four factors are active. Design 4 is best when the full five-factor model is considered but remains ranked in the top three for all weight combinations.

Figure 5 looks at the synthesized efficiency plot for Designs 1 and 4 to see how much lost efficiency exists for the designs across all of the weight combinations. Since only white and the lightest shade of gray are present in the plots, both Designs 1 and 4 are at least 90% efficient relative to the best available design at any weight combination of the criteria. Since the designs seem to be quite similar in performance for the primary criteria for this case, it can be helpful to look at Table 1 (and possibly A1 and A2) for information about other secondary criteria that might also be of interest. As mentioned in Section 2, some of the criteria were reduced to single number summaries for ease of comparison, but since this choice between Designs 1 and 4 involves more complex trade-offs, it can also be helpful to look at some graphical summaries to help reveal more details. If prediction throughout the design space is of interest, the FDS plot shown in Figure 6A can help provide additional details about SPV values. On the basis of this summary, Design

TABLE 4 Sample scenarios when top three and top only methods select different top designs

			Optimization criteria			Optimality only			Near-optimality (top three)		
Case	DF	Scale	1	2	3	Rank = 1 (opt.%)	Rank = 2 (opt.%)	Rank = 3 (opt.%)	Rank = 1 (opt.%)	Rank = 2 (opt.%)	Rank = 3 (opt.%)
1	Add	All	pwr^2_{MT}	$pwr^{2,p4}_{MT}$	$pwr^{2,p3}_{MT}$	1 (63.2)	4 (36.8)	NA	4 (100)	3 (100)	1 (78.4)
2	Add	All	pwr^2_M	AC_{MT}	$tr(AA')$	1 (64.5)	4 (35.5)	NA	4 (100)	3 (100)	1 (71.0)
3	Add	All	pwr^3_T	AC_{MT}	$tr(R'R)$	4 (42.9)	50 (38.1)	52 (14.7)	50 (61)	4 (54.1)	52 (45.9)
4	Add	All	pwr^2_M	AC_{MT}	A^{p3}	1 (63.2)	4 (36.8)	NA	4 (100)	3 (100)	1 (70)
5	Add	All	pwr^2_M	AC_{MT}	D^{p3}	1 (61.9)	4 (38.1)	NA	4 (100)	3 (100)	1 (68.4)
6	Add	All	pwr^2_M	AC_{MT}	I^{p3}	1 (52.4)	4 (47.6)	NA	4 (100)	3 (100)	1 (60.6)
7	Add	All	pwr^2_M	AC_{MT}	G^{p3}	1 (64.5)	4 (35.5)	NA	4 (100)	3 (100)	1 (71)
8	Multi	All	D	I	$tr(R'R)$	19 (42.9)	24 (37.2)	26 (10.4)	26 (57.6)	19 (54.6)	24 (49.4)
9	Multi	All	A	G	$tr(R'R)^{p3}$	4 (70.1)	19 (16)	24 (13.9)	19 (97)	4 (77.1)	34 (65.8)
10	Add	PF	I	I^{p4}	I^{p3}	1 (75.8)	4 (24.2)	NA	4 (100)	3 (100)	1 (97)

The superscripts $p3$ and $p4$ indicate the criteria based on the projected designs with three and four design factors, respectively.

Abbreviations: $AC_{M\times T}$, average correlation over pairs of main effect and two-factor interaction; AC_{MT}, average correlation over pair of effects from either main effects or two-factor interactions; AC_T, average correlation over pairs of two-factor interactions; DF, desirability functions; pwr^2_M, average power of main effects at $r = \frac{\delta}{\sigma} = 2$; pwr^2_{MT}, average power of main effects and two-factor interactions; pwr^2_T, average power of two-factor interactions.

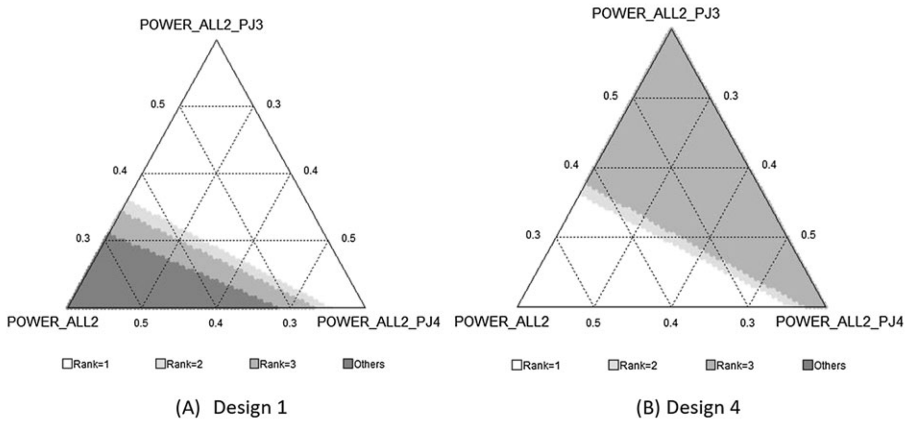

FIGURE 4 The design rank plot (DRP) for comparing Design 1 selected by considering only top design vs. Design 4 selected by considering top three designs for case 1 in Table 4 based on the following three criteria: pwr_{MT} (the average power for all main effects and two-factor interactions for the five-factor model), pwr_{MT}^{p4} (average over five models projected down to four design factors), and pwr_{MT}^{p3} (average over 10 models projected down to 3 design factors)

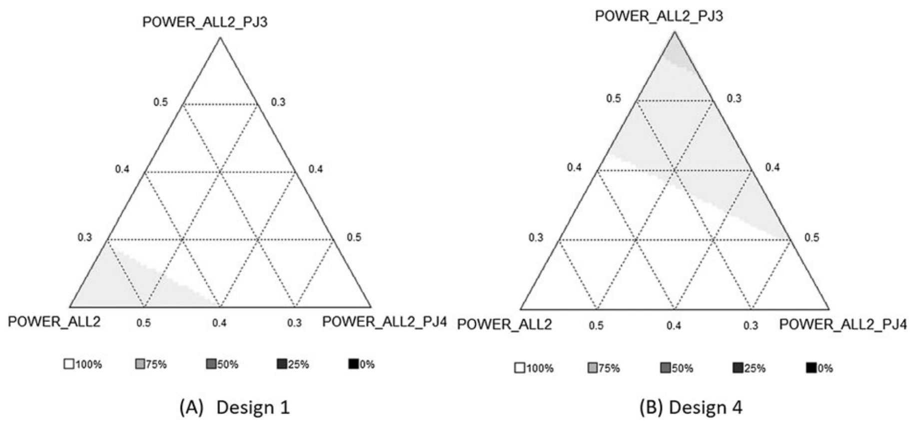

FIGURE 5 The synthesized efficiency plot (SEP) for comparing Design 1 selected by considering only top design vs. Design 4 selected by considering top three designs for case 1 in Table 4 based on the following three criteria: pwr_{MT}, pwr_{MT}^{p4}, and pwr_{MT}^{p3}

4 is superior (lower and flatter) throughout the region than Design 1. Similarly, if one of the secondary criteria of interest is the absolute correlation between terms in the model, the correlation color maps provide more detail than the averages reported in Table 1. In examining the two designs, we see that Design 1 has zero correlations between the main effects and all two-factor interactions while having more nonzero correlations between pairs of interactions. Design 4 has fewer overall nonzero correlations, but this includes some between main effects and two-factor interactions. Again, we see that choosing between the two designs involve trade-offs with no universal winner. The experimenter will need to carefully consider what to prioritize between competing objectives.

Consider case 9 in Table 4, a specialized case to illustrate a different pair of top designs. In this scenario, assume that a multiplicative DF with scaling based on all of the designs in Table 1 is of interest. The selected three primary criteria are *A*-efficiency, *G*-efficiency, and $\text{tr}(\boldsymbol{R}'\boldsymbol{R})$ for models based on three-factor models. For this scenario, the top design based on optimality is Design 4, which is top ranked for 70.1% of the weight combinations. However, if we are focused on near-optimality, then Design 19 is ranked in the top three for 97% of the weight combinations. Figure 7 shows the DRP for these two designs and highlights that Design 4 is best for weight combinations where *A*- and *G*-efficiencies are weighted more highly. Design 19 is top ranked for only 16% of weight combinations in a region where $\text{tr}(\boldsymbol{R}'\boldsymbol{R})$ is weighted about 40%-45%. When we examine the synthesized efficiency plots for these two designs in Figure 8, we see that the range of efficiencies across the different weight combinations varies considerably more than in previous cases. Here the worst efficiency can be quite small relative to the best available design, and so the trade-offs between the alternatives are more severe.

Figure 9 shows the mixture plot from the TopN-PFS add-in with the ranking of the top designs for different

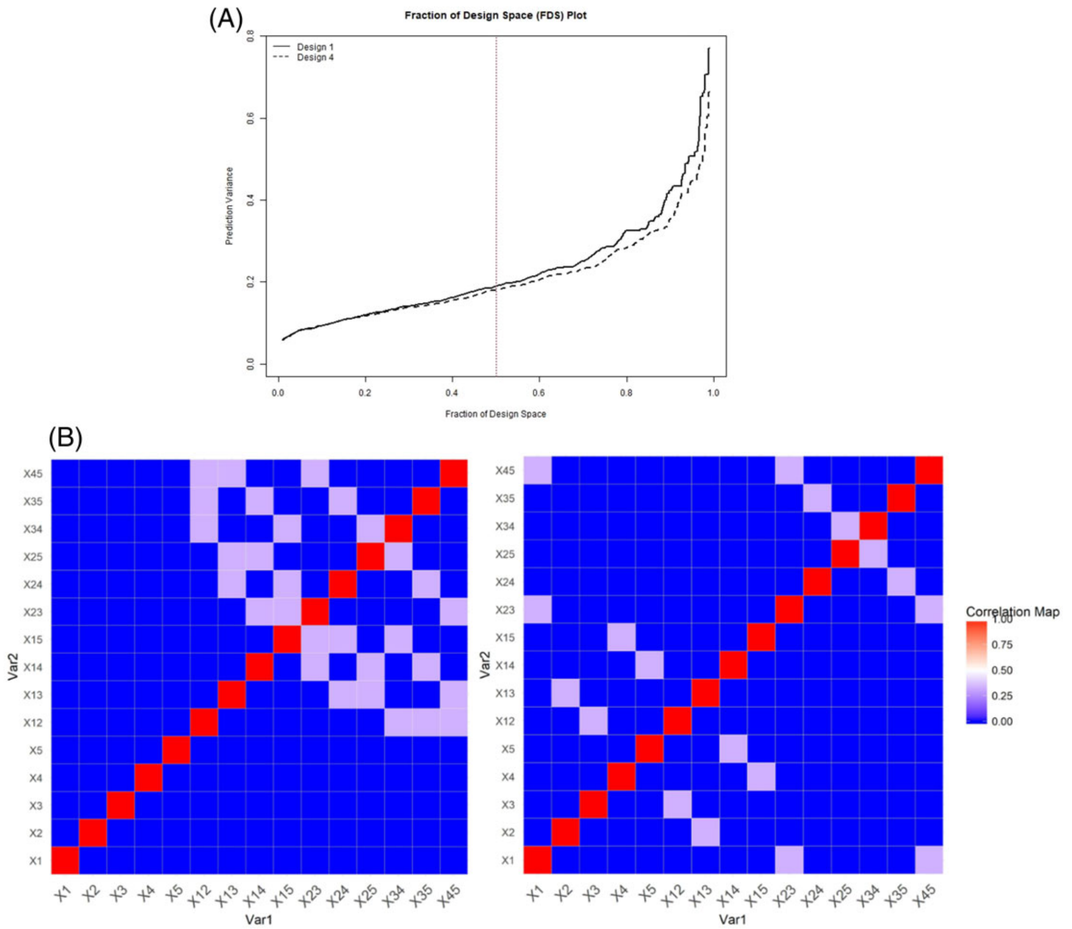

FIGURE 6 Other graphical summaries to support further design comparison among the selected top designs (Designs 1 & 4) based on using the criteria pwr_{MT}, pwr_{MT}^{p4}, and pwr_{MT}^{p3} [Colour figure can be viewed at wileyonlinelibrary.com]

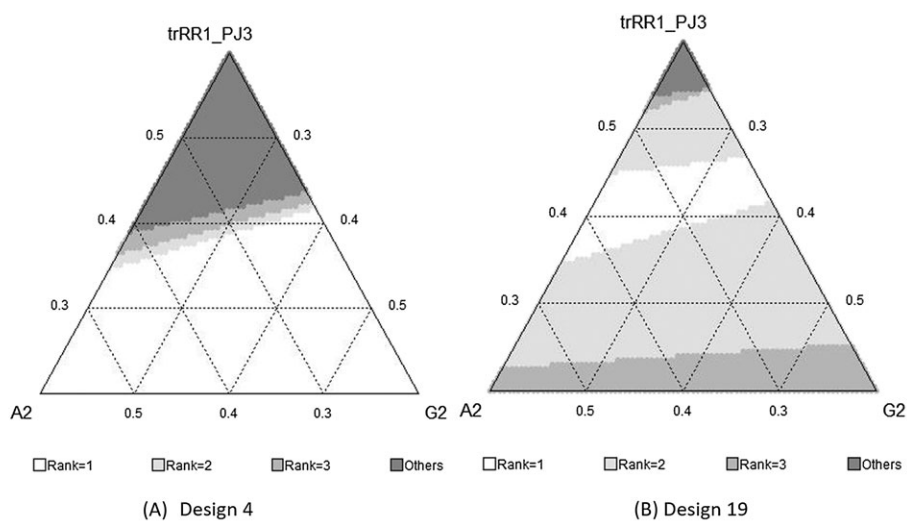

FIGURE 7 The design rank plot (DRP) for comparing Design 4 selected by considering only top design vs. Design 19 selected by considering top three designs based on the following three criteria (case 9 in Table 4): A-efficency, G-efficiency, and $\mathrm{tr}(R'R)^{p3}$ (the average $\mathrm{tr}(R'R)$ for 10 designs projected down to three factors)

weight combinations when the A-efficiency weight is fixed at 0.5. Results are shown for cases when the weights for the other two criteria are varied across the range of options from 0 to 0.5. For example, when the weights are $\left(w_A, w_G, w_{\mathrm{tr}(R'R)}\right) = (0.5, 0.3, 0.2)$, the top three designs are Design 4 (black), Design 2 (medium gray),

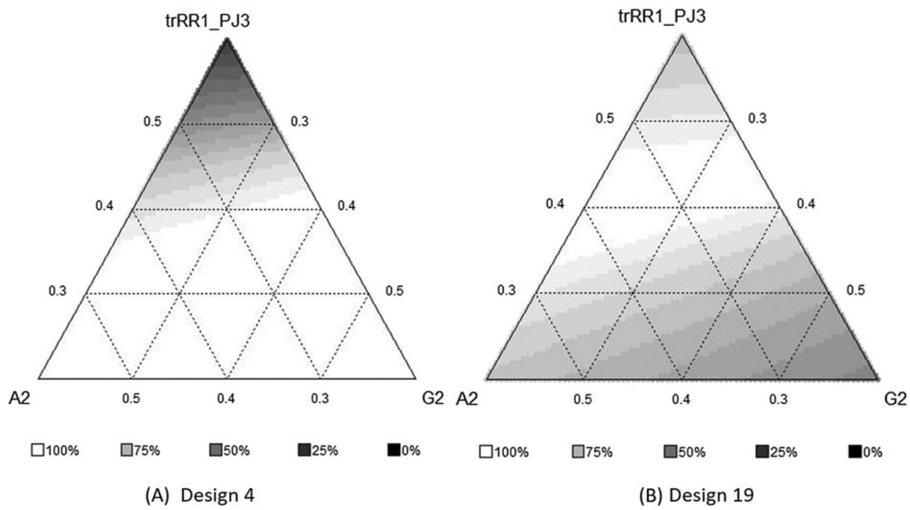

FIGURE 8 The synthesized efficiency plot (SEP) for comparing Design 4 selected by considering only top design vs. Design 19 selected by considering top three designs based on the following three criteria: *A*-efficency, *G*-efficiency, and $\text{tr}(R^{'}R)^{p3}$ (the average $\text{tr}(R^{'}R)$ for 10 designs projected down to three factors)

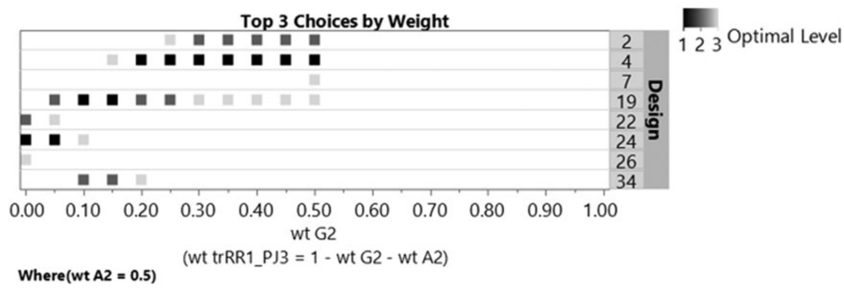

FIGURE 9 The mixture plot from the JMP add-in for showing the top three designs with the weight of A_2 fixed at 50%, and the weights for other criteria varying between 0 and 50%. Darker to lighter gray shades indicate high to low ranks

and Design 19 (lighter gray). This provides a complement to the information in Figures 7 and 8 for the two specific designs compared.

If we introduce secondary criteria to consider, Figure 10 shows the FDS and correlation color maps for the two designs. Again, Design 4 has superior performance with lower SPV values throughout the design region. The color map also shows Design 4 has fewer nonzero correlations between terms in the model and fewer between main and two-factor interaction terms. Although this case has rather specialized criteria that have been considered, it is helpful to examine how the described process for examining top designs for both optimality and near-optimality can lead to better understanding of their performance when different facets of the decision are considered.

With the R code provided, any set of three criteria selected of primary importance for the experimenter can be evaluated with quantification of the fraction where the designs are ranked first or in the Top *N*. In addition, the design rank and the synthesized efficiency plots enhance that information with more detail about which weight combinations are associated with higher performance for which criteria. The complete summary table with all 69 criteria values as well as the other supporting graphical summaries such as the FDS plot and correlation map add additional information to help break ties and choose between the most competitive designs.

In addition to the 24-run designs presented in detail here, there are other catalogs of relevant designs presented in Schoen et al (2010).[8]. In Table B1, we supply some matching criteria values for a subset of the designs from the catalog for five factors and 28 runs. Best values for each column are shown in bold. Again designs for which the *D*- and *A*-efficiency values were extremely poor have been removed for brevity. When similar sets of three criteria are selected for this catalog, there is greater consistency of results between the optimality and near-optimality rankings. This is perhaps explained by the larger design size relative to number of terms in the model to be estimated. As there is less pressure on the designs with more generous sample sizes, the trade-offs between designs become less extreme.

4 | R AND JMP TOOLS

This section describes the available R package for calculating all the criterion values for the catalog described

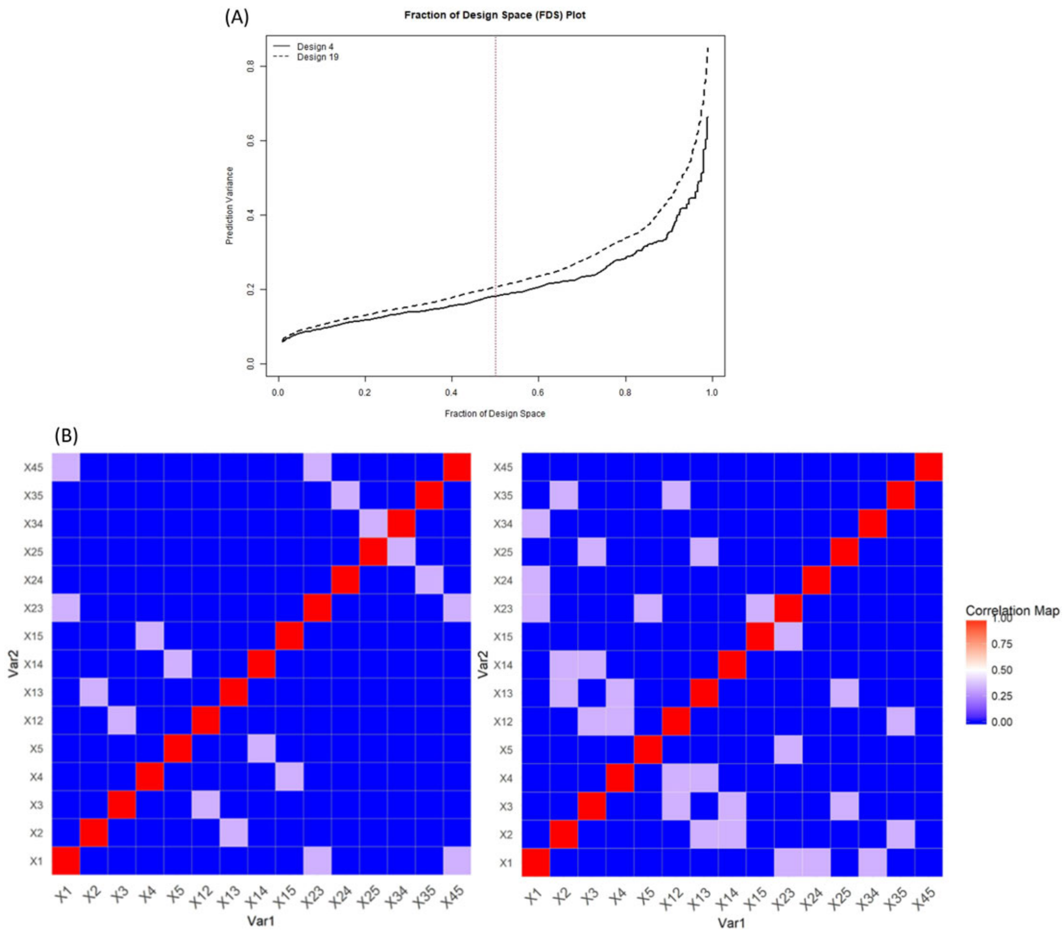

FIGURE 10 Other graphical summaries to support further design comparison among the selected top designs (Designs 4 & 19) based on the three criteria *A*- and *G*-efficiencies and $\mathrm{tr}(R'R)^{p3}$ [Colour figure can be viewed at wileyonlinelibrary.com]

and examining groups of criteria to identify the Top N ranked choices for any weight combination of any of the available criteria. In addition, it summarizes the functionality of the TopN-PFS JMP add-in.

4.1 | R package MCDS

The R package MCDS, available from the authors by request, contains functions for performing two main tasks. The first is design evaluation based on the large set of diverse design criteria discussed in Section 2 while the second task is design selection and comparison. This second task is performed based on user-selected subsets of three criteria using both the optimality (finding only the best choice) and near-optimality (considering those choices ranked in the Top N) approaches. The main design evaluation function calculates 34 design criteria for each candidate design in a user-provided list of design matrices with five factors. The criteria include all 23 criteria used for the 24-run design illustrated in Section 3 and additional 12 criteria that are generally of interest to designs with three or higher levels for each factor and considering second- or third-order models. Two additional design evaluation functions can evaluate the list of designs based on the average performance of smaller designs considering all subsets of fewer active factors (particularly for three or four active factors). A summary table (similar to Table 1) with up to 102 calculated design criteria values is automatically generated.

For the second task, the package considers three criteria as the primary quantitative criteria for design selection. This restriction is based on the consideration that including too many design criteria often results in too many choices remaining in the Reduce stage of DMRCS. This typically results in a list of choices that is hard to manage and selects designs with mediocre performance. The package offers the flexibility for the users to explore different combinations of the three chosen design criteria for understanding trade-offs and comparing solutions. After applying the optimality- and near-optimality–based design selection approaches, the package creates summary tables (similar to Tables 3 & 4) to show the top 3 choices with the corresponding weight regions to

TABLE 5 The table summarizes the selected best and worst criteria values for the 28-run designs

	Five-factor design		Projection to four-factor designs		Projection to three-factor designs	
Criteria	Best	Worst	Best	Worst	Best	Worst
D-efficiency	0.941	0.640	0.971	0.859	0.991	0.954
A-efficiency	0.883	0.185	0.942	0.711	0.982	0.910
I	0.156	0.668	0.243	0.317	0.351	0.378
G-efficiency	0.759	0.041	0.747	0.362	0.875	0.742
pwr_M^2	0.895	0.299	0.927	0.813	0.942	0.913
pwr_T^2	0.895	0.316	0.927	0.834	0.942	0.913
pwr_{MT}^2	0.895	0.323	0.927	0.831	0.942	0.913
AC_T	0.048	0.086	0.029	0.051	0	0
$AC_{M \times T}$	0.086	0.171	0.071	0.143	0.048	0.095
AC_{MT}	0.061	0.110	0.048	0.086	0.029	0.057
$\mathrm{tr}(AA')$	0.61	3.06	0.25	1.22	0.06	0.31
$\mathrm{tr}(R'R)$	194	263	134	161	75	82

Abbreviations: $AC_{M \times T}$, average correlation over pairs of main effect and two-factor interaction; AC_{MT}, average correlation over pair of effects from either main effects or two-factor interactions; AC_T, average correlation over pairs of two-factor interactions; pwr_M^2, average power of main effects at $r = \frac{\delta}{\sigma} = 2$; pwr_{MT}^2, average power of main effects and two-factor interactions; pwr_T^2, average power of two-factor interactions.

indicate robustness of performance using the two methods for all specified combinations of criteria.

The package also has four functions to generate four graphical summaries (design rank plot, synthesized efficiency plot, fraction of design space plot, and the correlation color map) to facilitate descriptive design comparisons between the identified top choices. The package also allows exploration of smaller weight regions as specified by lower bounds for individual criteria within the entire weight space when more focused user priorities[21] are desired. The details of all of the designs considered in the 24-run example illustrated in Section 3 with a list of different combinations of interested design criteria are also included in the package to illustrate the use of the main functions.

4.2 | JMP add-in, TopN-PFS

The JMP add-in implements the TopN-PFS algorithm described in Section 2. The add-in has a user-interface that allows the experimenter to customize the results based on the priorities of a given test. Given a table with enumerated criteria for each solution, the add-in identifies the solutions in the Top N layered PFs in addition to several graphical summaries to help guide the experimenters to a decision. The user first specifies which columns in the data table contain the criteria of interest in addition to an ID column. The user can then enter in required specifications for their scenario. These include the number of PF layers to extract, how to scale the criteria, the DF, and which graphical summaries to display.

All results are shown in a new window that identifies on which PF layer each solution lies in both a table and in a pairwise scatterplot matrix. Other graphical displays include the mixture plot, proportion plot, parallel plot, and synthesized efficiency plot. The mixture plot uses gray-scale coloring to show the Top N solutions across all evaluated weight combinations. The proportion plot is a stacked bar chart that indicates how frequently each solution is a Top N solution across all weight combinations of the criteria. The parallel plot allows the user to visually examine the trade-offs of one solution over another in terms of the scaled criteria values. Finally, the synthesized efficiency plot, similar to the mixture plot, shows the robustness of each solution across the weight combinations. Across each weight combination, scaled DF scores to the best solution at that weight combination are graphed for each solution. Dark blue boxes in this plot indicate the DF score for that solution is very similar to the optimal solution at that weight combination. Light blue boxes indicate poor performance relative to the optimal solution at that weight combination.

The TopN-PFS add-in is available on the JMP user community page (https://community.jmp.com/t5/JMP-Add-Ins/Top-N-Pareto-Front-Search-for-Structured-Decision-Making/ta-p/36527). A user-guide is included on this page. In addition, the add-in itself includes a help file in the opening dialog box that includes an example demonstrating the capabilities of the add-in.

5 | CONCLUSIONS

In this paper we have described an approach to consider near-optimal performance with improved robustness to broad user priorities as a complement to strictly optimal performance for more restricted prioritization of a set of design criteria suitable for a particular experiment. A first most important consideration in the process of selecting a design for a testing scenario is to think carefully about the priorities of the study and how to measure these aspects with quantitative measures. Section 2 outlines a number of different criteria that may be suitable for different experiments. Often for screening experiments, good estimation of the model parameters using *D*- or *A*-optimality is a primary consideration. However, this objective can be complemented with other characteristics of the design that provide protection against things going wrong or flexibility if the model is different from the one that was originally anticipated. Choosing the right set of objectives over which to optimize is a key to obtaining the best design for the experiment.

Once the criteria have been selected, then a PF approach can be used to objectively eliminate noncontending choices from further consideration. This helps keep the decision space manageable while focusing on the most promising contenders. The novel approach proposed in this paper is to value near-optimality with improved robustness to different user priorities as a complement to more localized strictly optimal performance. Since different people and organizations using the results from operational tests (or in other applications) often have divergent views of the importance of the different objectives of the test, selecting a method that allows for these different priorities to be considered and evaluated is beneficial. We show that there are situations where selecting a design with near-optimality for more robust combinations of the different criteria compared with selecting a local optimal design under more restricted conditions can be a good strategy to balance these divergent views when selecting the designed experiment to implement.

The method of considering how frequently designs are ranked first for different weights of a desirability function demonstrates one level of robustness. We extend this with identifying the top designs that are ranked in the Top *N* choices for the greatest fraction of weights of interest for the desirability function form chosen to demonstrate robustness based on considering near-optimality. We illustrate the method for a catalog of five-factor 24-run designs and show that different choices of sets of criteria to use for selecting the design make an important difference in the preferred design. Clearly there is no universal winner, and what is important for the experiment goals should be reflected in the choice of final design. Tables of criteria values for diverse objectives are provided as well as some summaries of the leading designs for selecting subsets of the criteria. A table of criteria values for a catalog of five-factor 28-run designs[8] is also provided to allow a practitioner considering this design scenario to conduct their own selection process (Table 5).

Note that the availability of an existing catalog of designs to explore does offer advantages of having deeper understanding of the possible choices across the design space of interest and a simplified design construction process (select from the list instead of search through a large design space). However, the proposed method of using the layered Pareto front approach to evaluate near-optimal designs for improving the robustness of selected designs to different user priorities can be adapted for the more general practice where designs are constructed from scratch for evaluation and selection. This requires development of a search algorithm for identifying the layered Pareto fronts based on the multiple criteria. Existing algorithms for searching the regular PFs such as the PAPE algorithm[2] and its variations[14], genetic algorithms[13], and NSGA-II algorithms[16,17] can be adapted with the similar design selection process outlined in Section 2.2 built in the search algorithm for seeking the layered PFs.

With the great flexibility offered by the proposed method to explore broader choices of design criteria and user priorities, it requires the experimenters to (a) be sufficiently DOE-informed, (b) understand what are the important characteristics to consider for design selection (see page 370 of Myers et al[10] for a useful list of design properties), (c) understand which characteristics are most relevant or important to their particular experiment, and (d) understand which criteria are useful for measuring those characteristics. We recommend all DOE decision-makers to make this important investment prior to using any design optimization and selection tools to avoid making naïve or undesirable decisions. The proposed method is advantageous on offering tools and procedures to explore different possibilities and uncovering the hidden trade-offs that are critical to select the best design for their particular experimental goals and supporting more informed and justifiable decisions.

Finally, a reviewer suggested that there may be other approaches than looking at ranks to combine the information about each design's performance. Perhaps a weighted sum or average of the raw metrics might be used.

ACKNOWLEDGMENTS

The authors thank the editor, the associate editor, and reviewers for their valuable comments and suggestions that led to significant improvements on this paper. This research was supported by the Office of the Secretary of Defense, Director of Operational Test and Evaluation (OSD DOT&E) and the Test Resource Management Center (TRMC) within the Science of Test research consortium.

DISCLAIMER

The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the US Government.

REFERENCES

1. Derringer G, Suich R. Simultaneous optimization of several response variables. *J Qual Technol*. 1980;12(4):214-219.
2. Lu L, Anderson-Cook CM, Robinson TJ. Optimization of designed experiments based on multiple criteria utilizing Pareto frontier. *Dent Tech*. 2011;53:353-365.
3. Lu L, Anderson-Cook CM. Rethinking the optimal response surface design for a first-order model with two-factor interactions, when protecting against curvature. *Qual Eng*. 2012;24(3):404-421.
4. Lu L, Anderson-Cook CM. Balancing multiple criteria incorporating cost using Pareto front optimization for split-plot designed experiments. *Qual Reliab Eng Int*. 2014;30(1):37-55.
5. Lu L, Anderson-Cook CM, Robinson TJ. A case study to demonstrate Pareto frontiers for selecting a best response surface design with simultaneously optimizing multiple criteria. *Appl Stoch Model Bus Ind*. 2012;28(3):206-221.
6. Burke SE, Anderson-Cook CM, Lu L, Montgomery DC. Prioritization of stockpile maintenance with layered Pareto fronts. *Qual Eng*. 2018. (in press)
7. Anderson-Cook CM, Lu L. Much-needed structure: a new 5-step decision-making process helps you evaluate, balance competing objectives. *Quality Progress*. 2015;48(10):42-50.
8. Schoen ED, Eendebak PT, Nguyen MVM. Complete enumeration of pure-level and mixed-level orthogonal arrays. *J Comb Des*. 2010;18:123-140.
9. Jones B, Schoen ED, Montgomery DC. A comparison of two-level designs to estimate all main effects and two-factor interactions. *Qual Eng*. 2016;28(4):369-380.
10. Myers RH, Montgomery DC, Anderson-Cook CM. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. 4th ed. New York: Wiley; 2016.
11. Zahran A, Anderson-Cook CM, Myers RH. Fraction of design space to assess the prediction capability of response surface designs. *J Qual Technol*. 2003;35(4):377-386.
12. Lu L, Robinson TJ, Anderson-Cook CM. A case study to select an optimal split-plot design for a mixture-process experiment based on multiple objectives. *Qual Eng*. 2014;26(4):424-439.
13. Lu L, Chapman JL, Anderson-Cook CM. A case study on selecting a best allocation of new data for improving the estimation precision of system and sub-system reliability using Pareto fronts. *Dent Tech*. 2013;55:473-487.
14. Lu L, Anderson-Cook CM. Adapting the hypervolume quality Indicator to quantify trade-offs and search efficiency for multiple criteria decision-making using Pareto fronts. *Qual Reliab Eng Int*. 2013;29(8):1117-1133.
15. Eschenauer H, Koski J, Osyczka AE. *Multicriteria Design Optimization: Procedures and Applications*. Berlin: Springer; 1990.
16. Deb K. *Multi-Objective Optimization using Evolutionary Algorithms*. New York: John Wiley & Sons; 2008.
17. Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans Evol Comput*. 2002;6(2):182-197.
18. Burke SE, Anderson-Cook CM, Lu L, Borror CM, Montgomery DC. Design of experiment selection using layered fronts *Los Alamos National Laboratory Technical Report*, LAUR-16-25643; 2016.
19. Hedayat AS, Sloane NJA, Stufken J. *Orthogonal Arrays: Theory and Applications*. New York: Springer; 1999.
20. JMP®. *Version* 13. Cary, NC: SAS Institute Inc.; 2017.
21. Lu L, Anderson-Cook CM, Lin D. Optimal designed experiments using a Pareto front search for focused preference of multiple objectives. *Comput Statist Data Anal*. 2014;71:1178-1192.

AUTHOR BIOGRAPHIES

Sarah E. Burke works for the Perduco Group as a statistician at the Scientific Test and Analysis Techniques Center of Excellence in Dayton, Ohio. She earned her doctorate in industrial engineering at Arizona State University. Her research interests include design of experiments, response surface methodology, multi-criteria decision making, and statistical engineering.

Lu Lu is an Assistant Professor of Statistics in the Department of Mathematics and Statistics at the University of South Florida in Tampa. She was a postdoctoral research associated in the Statistics Sciences Group at Los Alamos National Laboratory. She earned a doctorate in Statistics from Iowa State University in Ames, IA. Her research interests include reliability analysis, design of experiments, response surface methodology, survey sampling, multiple objective/response optimization.

Christine M. Anderson-Cook is a research scientist in the Statistical Sciences Group at Los Alamos National Laboratory. She is a Fellow of the ASA and ASQ. Her research areas include design of experiments, response surface methodology, reliability, and multiple criteria optimization.

Douglas C. Montgomery is Regents' Professor of industrial engineering and ASU Foundation Professor of Engineering at Arizona State University. His research and teaching interests are in industrial statistics. Professor Montgomery is an honorary member of the ASQ, a fellow of the ASA, a fellow of the RSS, a fellow of IIE, a member of the ISI, and an academician of the IAQ. He has received several teaching and research awards, including the Outstanding Doctoral Mentor award and the President's Award for Innovation from ASU, the Shewhart Medal, the Distinguished Service Medal, the Brumbaugh Award, the Lloyd Nelson Award, and the William G. Hunter Award from the ASQ, the George Box Medal from ENBIS, the Greenfield Medal from the Royal Statistical Society, and the Deming Lecture Award from the American Statistical Association. He is a former editor of the Journal of Quality Technology and one of the chief editors of Quality and Reliability Engineering International.

How to cite this article: Burke SE, Lu L, Anderson-Cook CM, Montgomery DC. Selecting a near-optimal design for multiple criteria with improved robustness to different user priorities. *Qual Reliab Engng Int*. 2018;1–27. https://doi.org/10.1002/qre.2413

APPENDIX

TABLE A1 Selected projection criteria values for Jones et al[9] catalog of five-factor designs with 24 runs averaged over five designs projected down to four design factors

Design	D^{p4}	A^{p4}	I^{p4}	G^{p4}	$pwr_M^{2,p4}$	$pwr_T^{2,p4}$	$pwr_{MT}^{2,p4}$	AC_T^{p4}	$AC_{M\times T}^{p4}$	AC_{MT}^{p4}	$tr(AA')^{p4}$	$tr(R'R)^{p4}$
1	**0.968**	**0.936**	**0.240**	**0.786**	**0.892**	0.855	**0.870**	0.067	**0.000**	**0.022**	**0.00**	144
2	**0.968**	**0.936**	0.244	**0.786**	0.870	**0.870**	**0.870**	**0.013**	0.033	**0.022**	0.27	138
3	**0.968**	**0.936**	0.242	**0.786**	0.881	0.862	**0.870**	0.040	0.017	**0.022**	0.13	141
4	**0.968**	**0.936**	0.244	**0.786**	0.870	**0.870**	**0.870**	**0.013**	0.033	**0.022**	0.27	138
6	0.935	0.868	0.260	0.691	0.864	0.814	0.834	0.040	0.033	0.031	0.27	138
7	0.954	0.905	0.250	0.707	0.866	0.849	0.856	0.040	0.033	0.031	0.27	138
8	0.954	0.905	0.253	0.707	0.854	0.857	0.856	**0.013**	0.050	0.031	0.40	134
9	0.920	0.837	0.268	0.613	0.849	0.801	0.820	0.040	0.050	0.040	0.40	134
10	0.939	0.874	0.259	0.629	0.850	0.836	0.842	0.040	0.050	0.040	0.40	134
11	0.924	0.843	0.267	0.550	0.834	0.823	0.827	0.040	0.067	0.049	0.53	131
12	0.909	0.811	0.273	0.471	0.830	0.802	0.813	0.067	0.067	0.058	0.53	131
13	0.924	0.843	0.267	0.550	0.834	0.823	0.827	0.040	0.067	0.049	0.53	131
15	0.894	0.780	0.281	0.393	0.815	0.789	0.799	0.067	0.083	0.067	0.67	128
16	0.954	0.905	0.253	0.707	0.852	0.858	0.856	**0.013**	0.050	0.031	0.40	134
17	0.909	0.811	0.276	0.471	0.817	0.811	0.813	0.040	0.083	0.058	0.67	128
18	0.905	0.805	0.277	0.534	0.831	0.789	0.806	0.040	0.067	0.049	0.53	131
19	0.939	0.874	0.261	0.629	0.837	0.845	0.842	**0.013**	0.067	0.040	0.53	131
21	0.924	0.843	0.267	0.550	0.832	0.824	0.827	0.040	0.067	0.049	0.53	131
22	0.909	0.811	0.279	0.471	0.801	0.821	0.813	**0.013**	0.100	0.058	0.80	125
24	0.909	0.811	0.279	0.471	0.799	0.823	0.813	**0.013**	0.100	0.058	0.80	125
25	0.909	0.811	0.276	0.471	0.817	0.811	0.813	0.040	0.083	0.058	0.67	128
26	0.924	0.843	0.270	0.550	0.819	0.833	0.827	**0.013**	0.083	0.049	0.67	128
27	0.909	0.811	0.276	0.471	0.815	0.812	0.813	0.040	0.083	0.058	0.67	128
28	0.894	0.780	0.287	0.393	0.784	0.809	0.799	**0.013**	0.117	0.067	0.93	122
31	0.894	0.780	0.284	0.393	0.799	0.799	0.799	0.040	0.100	0.067	0.80	125
34	0.938	0.876	0.264	0.702	0.833	0.841	0.838	**0.013**	0.067	0.040	0.53	131
35	0.878	0.752	0.299	0.388	0.763	0.794	0.781	**0.013**	0.133	0.076	1.07	**118**
36	0.893	0.783	0.287	0.466	0.796	0.795	0.795	0.040	0.100	0.067	0.80	125
40	0.878	0.752	0.296	0.388	0.778	0.783	0.781	0.040	0.117	0.076	0.93	122
43	0.902	0.800	0.290	0.597	0.799	0.798	0.798	0.040	0.050	0.040	0.67	128
47	0.902	0.800	0.293	0.597	0.785	0.807	0.798	**0.013**	0.067	0.040	0.80	125
48	0.887	0.768	0.299	0.519	0.781	0.787	0.784	0.040	0.067	0.049	0.80	125
49	0.872	0.737	0.311	0.440	0.749	0.785	0.770	**0.013**	0.100	0.058	1.07	**118**
50	0.902	0.800	0.294	0.597	0.782	0.810	0.798	**0.013**	0.067	0.040	0.80	125
51	0.887	0.768	0.303	0.519	0.766	0.796	0.784	**0.013**	0.083	0.049	0.93	122
52	0.887	0.768	0.303	0.519	0.764	0.798	0.784	**0.013**	0.083	0.049	0.93	122

The best criteria values are highlighted in boldface. The superscript *p*4 indicates the criteria based on the projected designs with four design factors.

Abbreviations: $AC_{M\times T}$, average correlation over pairs of main effect and two-factor interaction; AC_{MT}, average correlation over pair of effects from either main effects or two-factor interactions; AC_T, average correlation over pairs of two-factor interactions; pwr_M^2, average power of main effects at $r = \frac{\delta}{\sigma} = 2$; pwr_{MT}^2, average power of main effects and two-factor interactions; pwr_T^2, average power of two-factor interactions.

TABLE A2 Selected projection criteria values for Jones et al[9] catalog of five-factor designs with 24 runs averaged over 10 designs projected down to three-design factors

Design	D^{p3}	A^{p3}	I^{p3}	G^{p3}	$pwr_{M}^{2,p3}$	$pwr_{T}^{2,p3}$	$pwr_{MT}^{2,p3}$	AC_{T}^{p3}	$AC_{M\times T}^{p3}$	AC_{MT}^{p3}	$tr(AA')^{p3}$	$tr(R'R)^{p3}$
1	**1.000**	**1.000**	**0.346**	**1.000**	**0.904**	**0.904**	**0.904**	**0.000**	**0.000**	**0.000**	**0.00**	72
2	0.990	0.981	0.352	0.940	0.897	0.897	0.897	**0.000**	0.022	0.013	0.07	70
3	0.995	0.990	0.349	0.970	0.900	0.900	0.900	**0.000**	0.011	0.007	0.03	71
4	0.990	0.981	0.352	0.940	0.897	0.897	0.897	**0.000**	0.022	0.013	0.07	70
6	0.990	0.981	0.352	0.940	0.897	0.897	0.897	**0.000**	0.022	0.013	0.07	70
7	0.990	0.981	0.352	0.940	0.897	0.897	0.897	**0.000**	0.022	0.013	0.07	70
8	0.985	0.971	0.355	0.910	0.893	0.893	0.893	**0.000**	0.033	0.020	0.10	70
9	0.985	0.971	0.355	0.910	0.893	0.893	0.893	**0.000**	0.033	0.020	0.10	70
10	0.985	0.971	0.355	0.910	0.893	0.893	0.893	**0.000**	0.033	0.020	0.10	70
11	0.980	0.961	0.357	0.880	0.889	0.889	0.889	**0.000**	0.044	0.027	0.13	69
12	0.980	0.961	0.357	0.880	0.889	0.889	0.889	**0.000**	0.044	0.027	0.13	69
13	0.980	0.961	0.357	0.880	0.889	0.889	0.889	**0.000**	0.044	0.027	0.13	69
15	0.975	0.952	0.360	0.850	0.886	0.886	0.886	**0.000**	0.056	0.033	0.17	68
16	0.985	0.971	0.355	0.910	0.893	0.893	0.893	**0.000**	0.033	0.020	0.10	70
17	0.975	0.952	0.360	0.850	0.886	0.886	0.886	**0.000**	0.056	0.033	0.17	68
18	0.980	0.961	0.357	0.880	0.889	0.889	0.889	**0.000**	0.044	0.027	0.13	69
19	0.980	0.961	0.357	0.880	0.889	0.889	0.889	**0.000**	0.044	0.027	0.13	69
21	0.980	0.961	0.357	0.880	0.889	0.889	0.889	**0.000**	0.044	0.027	0.13	69
22	0.970	0.942	0.363	0.820	0.882	0.882	0.882	**0.000**	0.067	0.040	0.20	67
24	0.970	0.942	0.363	0.820	0.882	0.882	0.882	**0.000**	0.067	0.040	0.20	67
25	0.975	0.952	0.360	0.850	0.886	0.886	0.886	**0.000**	0.056	0.033	0.17	68
26	0.975	0.952	0.360	0.850	0.886	0.886	0.886	**0.000**	0.056	0.033	0.17	68
27	0.975	0.952	0.360	0.850	0.886	0.886	0.886	**0.000**	0.056	0.033	0.17	68
28	0.966	0.932	0.366	0.790	0.879	0.879	0.879	**0.000**	0.078	0.047	0.23	**66**
31	0.970	0.942	0.363	0.820	0.882	0.882	0.882	**0.000**	0.067	0.040	0.20	67
34	0.980	0.961	0.357	0.880	0.889	0.889	0.889	**0.000**	0.044	0.027	0.13	69
35	0.961	0.923	0.368	0.760	0.875	0.875	0.875	**0.000**	0.089	0.053	0.27	**66**
36	0.970	0.942	0.363	0.820	0.882	0.882	0.882	**0.000**	0.067	0.040	0.20	67
40	0.966	0.932	0.366	0.790	0.879	0.879	0.879	**0.000**	0.078	0.047	0.23	**66**
43	0.973	0.950	0.367	0.907	0.878	0.878	0.878	**0.000**	0.033	0.020	0.17	68
47	0.968	0.940	0.370	0.877	0.874	0.874	0.874	**0.000**	0.044	0.027	0.20	67
48	0.968	0.940	0.370	0.877	0.874	0.874	0.874	**0.000**	0.044	0.027	0.20	67
49	0.958	0.921	0.375	0.817	0.867	0.867	0.867	**0.000**	0.067	0.040	0.27	**66**
50	0.968	0.940	0.370	0.877	0.874	0.874	0.874	**0.000**	0.044	0.027	0.20	67
51	0.963	0.930	0.372	0.847	0.871	0.871	0.871	**0.000**	0.056	0.033	0.23	**66**
52	0.963	0.930	0.372	0.847	0.871	0.871	0.871	**0.000**	0.056	0.033	0.23	**66**

The best criteria values are highlighted in boldface. The superscript *p*3 indicates the criteria based on the projected designs with three design factors.

Abbreviations: $AC_{M\times T}$, average correlation over pairs of main effect and two-factor interaction; AC_{MT}, average correlation over pair of effects from either main effects or two-factor interactions; AC_T, average correlation over pairs of two-factor interactions; pwr_M^2, average power of main effects at $r = \frac{\delta}{\sigma} = 2$; pwr_{MT}^2, average power of main effects and two-factor interactions; pwr_T^2, average power of two-factor interactions.

TABLE B1 Selected criteria values for Jones et al[9] catalog of five-factor designs with 28 runs

Design	*D*-efficiency	*A*-efficiency	*I*	*G*-efficiency	pwr^2_M	pwr^2_T	pwr^2_{MT}	AC_T	$AC_{M \times T}$	AC_{MT}	$\mathrm{tr}(AA')$	$\mathrm{tr}(R'R)$
1	**0.941**	**0.883**	**0.156**	**0.759**	**0.895**	**0.895**	**0.895**	**0.048**	**0.086**	**0.061**	**0.61**	263
2	0.896	0.779	0.170	0.451	0.884	0.837	0.852	0.067	**0.086**	0.069	**0.61**	263
3	0.937	0.872	0.158	0.492	0.891	0.891	0.891	**0.048**	**0.086**	**0.061**	**0.61**	263
4	0.850	0.669	0.188	0.289	0.873	0.758	0.796	0.086	**0.086**	0.078	**0.61**	263
5	0.897	0.782	0.169	0.398	0.884	0.839	0.854	0.067	**0.086**	0.069	**0.61**	263
6	0.853	0.686	0.184	0.338	0.877	0.770	0.806	0.086	**0.086**	0.078	**0.61**	263
7	0.935	0.867	0.159	0.577	0.889	0.889	0.889	**0.048**	**0.086**	**0.061**	**0.61**	263
8	0.896	0.770	0.171	0.270	0.882	0.831	0.848	0.067	**0.086**	0.069	**0.61**	263
9	0.893	0.760	0.173	0.362	0.876	0.827	0.844	0.067	**0.086**	0.069	**0.61**	263
10	0.941	0.881	0.156	0.634	0.894	0.894	0.894	**0.048**	**0.086**	**0.061**	**0.61**	263
11	0.852	0.678	0.186	0.324	0.873	0.764	0.801	0.086	**0.086**	0.078	**0.61**	263
12	0.941	0.881	0.156	0.634	0.894	0.894	0.894	**0.048**	**0.086**	**0.061**	**0.61**	263
13	0.791	0.542	0.216	0.292	0.866	0.664	0.732	0.086	**0.086**	0.078	**0.61**	263
14	0.794	0.551	0.213	0.263	0.871	0.669	0.736	0.086	**0.086**	0.078	**0.61**	263
15	0.788	0.531	0.219	0.239	0.864	0.653	0.723	0.086	**0.086**	0.078	**0.61**	263
16	0.896	0.770	0.178	0.270	0.833	0.856	0.848	**0.048**	0.103	0.069	1.10	249
17	0.853	0.676	0.194	0.283	0.817	0.789	0.798	0.067	0.103	0.078	1.10	249
18	0.858	0.702	0.187	0.319	0.833	0.804	0.814	0.067	0.103	0.078	1.10	249
19	0.827	0.664	0.200	0.456	0.791	0.791	0.791	0.067	0.120	0.086	1.59	235
20	0.858	0.702	0.195	0.319	0.783	0.829	0.814	**0.048**	0.120	0.078	1.59	235
21	0.897	0.782	0.175	0.398	0.840	0.861	0.854	**0.048**	0.103	0.069	1.10	249
22	0.893	0.760	0.180	0.362	0.827	0.852	0.844	**0.048**	0.103	0.069	1.10	249
23	0.860	0.716	0.185	0.448	0.836	0.813	0.821	0.067	0.103	0.078	1.10	249
24	0.860	0.716	0.191	0.448	0.791	0.836	0.821	**0.048**	0.120	0.078	1.59	235
25	0.752	0.475	0.247	0.234	0.788	0.622	0.677	0.086	0.103	0.086	1.10	249
26	0.807	0.573	0.218	0.256	0.793	0.702	0.733	0.086	0.103	0.086	1.10	249
27	0.896	0.779	0.176	0.451	0.837	0.860	0.852	**0.048**	0.103	0.069	1.10	249
28	0.852	0.668	0.196	0.323	0.813	0.788	0.796	0.067	0.103	0.078	1.10	249
29	0.852	0.668	0.204	0.323	0.763	0.813	0.796	**0.048**	0.120	0.078	1.59	235
30	0.897	0.782	0.175	0.398	0.837	0.862	0.854	**0.048**	0.103	0.069	1.10	249
31	0.858	0.702	0.189	0.319	0.826	0.808	0.814	0.067	0.103	0.078	1.10	249
32	0.896	0.770	0.179	0.270	0.827	0.859	0.848	**0.048**	0.103	0.069	1.10	249
33	0.856	0.679	0.201	0.279	0.762	0.819	0.800	**0.048**	0.120	0.078	1.59	235
34	0.853	0.676	0.202	0.283	0.760	0.817	0.798	**0.048**	0.120	0.078	1.59	235
35	0.821	0.629	0.209	0.302	0.768	0.768	0.768	0.067	0.120	0.086	1.59	235
36	0.856	0.679	0.193	0.279	0.819	0.791	0.800	0.067	0.103	0.078	1.10	249
37	0.798	0.512	0.240	0.159	0.755	0.656	0.689	0.086	0.103	0.086	1.10	249
40	0.706	0.372	0.313	0.168	0.668	0.546	0.586	0.086	0.120	0.094	1.59	235
41	0.752	0.475	0.246	0.234	0.792	0.620	0.677	0.086	0.103	0.086	1.10	249
42	0.797	0.555	0.223	0.202	0.792	0.692	0.725	0.086	0.103	0.086	1.10	249
43	0.798	0.512	0.250	0.159	0.698	0.684	0.689	0.067	0.120	0.086	1.59	235

(Continues)

TABLE B1 (Continued)

Design	*D*-efficiency	*A*-efficiency	*I*	*G*-efficiency	pwr_M^2	pwr_T^2	pwr_{MT}^2	AC_T	$AC_{M\times T}$	AC_{MT}	tr(AA')	tr($R'R$)
44	0.850	0.669	0.195	0.289	0.817	0.786	0.796	0.067	0.103	0.078	1.10	249
45	0.853	0.686	0.192	0.338	0.822	0.797	0.806	0.067	0.103	0.078	1.10	249
46	0.754	0.470	0.263	0.220	0.700	0.630	0.653	0.086	0.120	0.094	1.59	235
47	0.805	0.581	0.215	0.279	0.804	0.709	0.741	0.086	0.103	0.086	1.10	249
48	0.806	0.581	0.215	0.283	0.801	0.711	0.741	0.086	0.103	0.086	1.10	249
49	0.853	0.686	0.191	0.338	0.826	0.795	0.806	0.067	0.103	0.078	1.10	249
50	0.753	0.443	0.279	0.171	0.680	0.619	0.640	0.086	0.120	0.094	1.59	235
51	0.750	0.464	0.263	0.166	0.706	0.618	0.647	0.086	0.120	0.094	1.59	235
52	0.722	0.286	0.413	0.073	0.532	0.481	0.498	0.086	0.120	0.094	1.59	235
53	0.798	0.550	0.225	0.244	0.784	0.689	0.721	0.086	0.103	0.086	1.10	249
54	0.807	0.573	0.226	0.256	0.738	0.730	0.733	0.067	0.120	0.086	1.59	235
55	0.852	0.678	0.193	0.324	0.820	0.791	0.801	0.067	0.103	0.078	1.10	249
56	0.753	0.443	0.279	0.171	0.674	0.622	0.640	0.086	0.120	0.094	1.59	235
57	0.752	0.432	0.285	0.179	0.662	0.606	0.624	0.086	0.120	0.094	1.59	235
58	0.772	0.460	0.260	0.144	0.742	0.608	0.653	0.086	0.103	0.086	1.10	249
59	0.750	0.464	0.264	0.166	0.703	0.620	0.647	0.086	0.120	0.094	1.59	235
60	0.753	0.443	0.291	0.171	0.629	0.645	0.640	0.067	0.137	0.094	2.08	222
61	0.738	0.394	0.309	0.181	0.640	0.609	0.619	0.086	0.120	0.094	1.59	235
62	0.853	0.686	0.199	0.338	0.770	0.824	0.806	**0.048**	0.120	0.078	1.59	235
63	0.806	0.581	0.225	0.283	0.736	0.744	0.741	0.067	0.120	0.086	1.59	235
64	0.806	0.581	0.235	0.283	0.686	0.769	0.741	**0.048**	0.137	0.086	2.08	222
65	0.853	0.686	0.200	0.338	0.769	0.824	0.806	**0.048**	0.120	0.078	1.59	235
66	0.754	0.470	0.265	0.220	0.685	0.637	0.653	0.086	0.120	0.094	1.59	235
67	0.657	0.185	0.647	0.041	0.337	0.316	0.323	0.086	0.137	0.102	2.08	222
68	0.738	0.394	0.343	0.181	0.577	0.640	0.619	**0.048**	0.154	0.094	2.57	208
69	0.666	0.254	0.468	0.079	0.445	0.410	0.421	0.086	0.137	0.102	2.08	222
70	0.850	0.669	0.204	0.289	0.761	0.813	0.796	**0.048**	0.120	0.078	1.59	235
71	0.805	0.581	0.226	0.279	0.734	0.744	0.741	0.067	0.120	0.086	1.59	235
72	0.798	0.512	0.253	0.159	0.682	0.692	0.689	0.067	0.120	0.086	1.59	235
73	0.850	0.669	0.205	0.289	0.754	0.817	0.796	**0.048**	0.120	0.078	1.59	235
74	0.798	0.512	0.265	0.159	0.618	0.724	0.689	**0.048**	0.137	0.086	2.08	222
75	0.656	0.242	0.497	0.096	0.417	0.389	0.398	0.086	0.137	0.102	2.08	222
76	0.805	0.581	0.234	0.279	0.685	0.769	0.741	**0.048**	0.137	0.086	2.08	222
77	0.656	0.242	0.552	0.096	0.308	0.444	0.398	**0.048**	0.171	0.102	3.06	**194**
78	0.797	0.555	0.246	0.202	0.658	0.759	0.725	**0.048**	0.137	0.086	2.08	222
79	0.752	0.432	0.298	0.179	0.606	0.633	0.624	0.067	0.137	0.094	2.08	222
80	0.852	0.678	0.201	0.324	0.764	0.819	0.801	0.048	0.120	0.078	1.59	235
81	0.754	0.470	0.287	0.220	0.587	0.686	0.653	**0.048**	0.154	0.094	2.57	208
82	0.754	0.470	0.277	0.220	0.631	0.664	0.653	0.067	0.137	0.094	2.08	222
83	0.657	0.185	0.668	0.041	0.299	0.335	0.323	0.067	0.154	0.102	2.57	208

(Continues)

TABLE B1 (Continued)

Design	*D*-efficiency	*A*-efficiency	*I*	*G*-efficiency	pwr^2_M	pwr^2_T	pwr^2_{MT}	AC_T	$AC_{M\times T}$	AC_{MT}	$tr(AA')$	$tr(R'R)$
84	0.657	0.185	0.651	0.041	0.333	0.318	0.323	0.086	0.137	0.102	2.08	222
85	0.752	0.432	0.299	0.179	0.605	0.634	0.624	0.067	0.137	0.094	2.08	222
87	0.807	0.573	0.236	0.256	0.678	0.760	0.733	**0.048**	0.137	0.086	2.08	222
88	0.753	0.443	0.293	0.171	0.615	0.652	0.640	0.067	0.137	0.094	2.08	222
89	0.711	0.232	0.533	0.052	0.379	0.400	0.393	0.067	0.137	0.094	2.08	222
90	0.807	0.573	0.229	0.256	0.724	0.737	0.733	0.067	0.120	0.086	1.59	235
91	0.711	0.232	0.518	0.052	0.422	0.379	0.393	0.086	0.120	0.094	1.59	235
93	0.753	0.443	0.293	0.171	0.620	0.649	0.640	0.067	0.137	0.094	2.08	222
97	0.722	0.286	0.434	0.073	0.500	0.497	0.498	0.067	0.137	0.094	2.08	222
98	0.798	0.550	0.236	0.244	0.714	0.724	0.721	0.067	0.120	0.086	1.59	235
99	0.666	0.254	0.481	0.079	0.436	0.414	0.421	0.086	0.137	0.102	2.08	222
100	0.798	0.550	0.247	0.244	0.663	0.749	0.721	**0.048**	0.137	0.086	2.08	222
101	0.666	0.254	0.503	0.079	0.383	0.440	0.421	0.067	0.154	0.102	2.57	208
102	0.722	0.286	0.448	0.073	0.462	0.516	0.498	0.067	0.137	0.094	2.08	222
104	0.750	0.464	0.291	0.166	0.575	0.683	0.647	**0.048**	0.154	0.094	2.57	208
105	0.750	0.464	0.285	0.166	0.597	0.673	0.647	0.067	0.137	0.094	2.08	222
106	0.772	0.460	0.295	0.144	0.564	0.697	0.653	**0.048**	0.137	0.086	2.08	222
109	0.640	0.314	0.395	0.147	0.471	0.471	0.471	0.086	0.154	0.110	2.57	208
110	0.791	0.542	0.251	0.292	0.664	0.765	0.732	**0.048**	0.120	0.078	2.08	222
112	0.794	0.551	0.247	0.263	0.669	0.770	0.736	**0.048**	0.120	0.078	2.08	222
116	0.752	0.475	0.286	0.234	0.606	0.713	0.677	**0.048**	0.137	0.086	2.57	208
118	0.706	0.372	0.350	0.168	0.546	0.607	0.586	0.067	0.137	0.094	2.57	208
120	0.752	0.475	0.276	0.234	0.635	0.698	0.677	0.067	0.120	0.086	2.08	222
123	0.788	0.531	0.256	0.239	0.653	0.758	0.723	**0.048**	0.120	0.078	2.08	222

The best criteria values are highlighted in boldface.

Abbreviations: $AC_{M\times T}$, average correlation over pairs of main effect and two-factor interaction; AC_{MT}, average correlation over pair of effects from either main effects or two-factor interactions; AC_T, average correlation over pairs of two-factor interactions; pwr^2_M, average power of main effects at $r = \frac{\delta}{\sigma} = 2$; pwr^2_{MT}, average power of main effects and two-factor interactions; pwr^2_T, average power of two-factor interactions.