Quality Engineering
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lqen20

A Case Study to Select an Optimal Split-Plot Design for a Mixture-Process Experiment Based on Multiple Objectives

Lu Lu[a], Timothy J. Robinson[b] & Christine M. Anderson-Cook[c]
[a] Department of Mathematics and Statistics, University of South Florida, Tampa, Florida
[b] WWAMI Medical Education Program and the Department of Statistics, University of Wyoming, Laramie, Wyoming
[c] Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, New Mexico
Published online: 20 Aug 2014.

To cite this article: Lu Lu, Timothy J. Robinson & Christine M. Anderson-Cook (2014) A Case Study to Select an Optimal Split-Plot Design for a Mixture-Process Experiment Based on Multiple Objectives, Quality Engineering, 26:4, 424-439, DOI: 10.1080/08982112.2014.887102

To link to this article: http://dx.doi.org/10.1080/08982112.2014.887102

Quality Engineering, 26:424–439, 2014

ISSN: 0898-2112 print / 1532-4222 online
DOI: 10.1080/08982112.2014.887102

Taylor & Francis
Taylor & Francis Group

A Case Study to Select an Optimal Split-Plot Design for a Mixture-Process Experiment Based on Multiple Objectives

Lu Lu[1],
Timothy J. Robinson[2],
Christine M. Anderson-Cook[3]
[1]Department of Mathematics and Statistics, University of South Florida, Tampa, Florida
[2]WWAMI Medical Education Program and the Department of Statistics, University of Wyoming, Laramie, Wyoming
[3]Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, New Mexico

ABSTRACT With increasingly constrained budgets, it is now becoming more desirable to get more information from each experiment and to have an intentional strategy for selecting designs for split-plot experiments that balance multiple competing objectives. Lu and Anderson-Cook (2014) developed a decision-making process for selecting an optimal split-plot design (SPD) for flexible objectives/criteria based on a Pareto front. The method allows exploration of all contending non-inferior choices with their trade-offs to enable an informed and justifiable decision based on understanding the potential impact of subjective aspects. This article considers a case study of a mixture-process experiment that seeks an SPD with a good balance of precise model coefficient estimates as measured by D-efficiency and low experimental cost, which is a function of both the time required to run the experiment as well as the financial cost. The D-efficiency is a function of the whole plot-to-subplot error variance ratio, a quantity that is typically not known a priori when the choice of a design must be made. The Pareto front approach is applied and graphical tools are used to quantify the trade-offs between criteria and robustness of design performance to different user-selected preferences for the criteria. A substantially different pattern of design performance robustness to the uncertainty of the specified variance ratio is demonstrated compared to non-mixture experiments.

KEYWORDS computer-generated designs, restricted randomization, multiple design criteria, design optimization, D-optimality, cost, Pareto front, point exchange search algorithm

INTRODUCTION

We consider a film manufacturing example in which the experimenter wishes to examine the relationship between the response (a measure of film quality) and five input factors: three whole plot/mixture factors (X_1, X_2, X_3) and two subplot/process factors (P_1, P_2). In producing the film, a combination of ingredients (the mixture factors) is mixed to form a roll. The roll is then cut into pieces and the subplot factor combinations are randomly applied to the pieces. Considerable time is needed in the formulation of each whole plot treatment

Address correspondence to Christine M. Anderson-Cook, Statistical Sciences Group, Los Alamos National Laboratory, P.O. Box 1663, MS F600, Los Alamos, NM 87545. E-mail: c-and-cook@lanl.gov

combination (i.e., the mixture combinations) and thus the mixture variables are treated as hard-to-change. Measuring film quality is expensive and the measurement process represents the majority of experimental cost. The experimenter wishes to balance both good model parameter estimation and reasonable cost when choosing an optimal split-plot design (SPD) for evaluating a 12-term Scheffé mixture model with three linear whole plot/mixture main effects, three two-factor interactions among the mixture variables and the six two-factor, mixture-by-process interactions of the form

$$y = \sum_{i=1}^{3} \beta_i X_i + \sum_{i<j}^{3} \beta_{ij} X_i X_j + \delta_{WP} + \sum_{i=1}^{3} \sum_{k=1}^{2} \alpha_{ik} X_i P_k + \varepsilon.$$

Note that because of the constraint that the sum of the ingredients must equal one, the process main effect terms are not included to ensure an estimable model.

The experimenter identified a feasible candidate set of seven mixture locations and four process locations as shown in Figure 1. Because the model contains terms for main effects and two-factor interactions for the whole plot mixture factors, levels [0, 0.5, 1] were considered as well as the overall centroid of the mixture space. For the subplot process factors, only two-way interactions with the whole plot factors are included in the model and hence high and low levels [+1 and −1] were deemed sufficient for the candidate set.

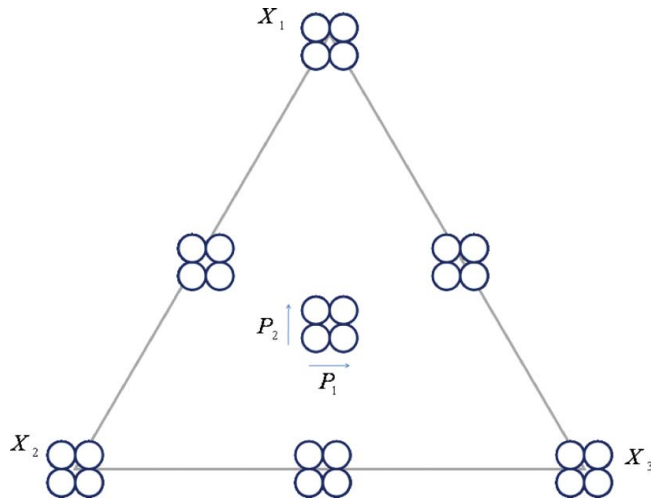

FIGURE 1 **Candidate space of 28 design locations. Clusters of four circles are centered at each of the seven potential whole plot mixture locations and each circle represents a +/− combination of the process factors, P_1 and P_2 (for example, lower left circle in the top cluster represents (X_1, X_2, X_3, P_1, P_2) = (1, 0, 0, −1, −1)).**

Preliminary data from a pilot study suggested that the whole plot–to-subplot error variance ratio was 5, but there was considerable uncertainty about this estimate. Consequently, subject matter experts provided a range of 1 to 10 as a bound for the true variance ratio. The D-criterion for SPDs (formally defined in the next section) depends upon the user specifying the whole plot–to-subplot error variance ratio d; hence, a D-optimal design for one value of d may not necessarily be D-optimal for another value. In this case, we consider the approach of evaluating the robustness of different designs across the range of anticipated d values. If designs can be identified to perform similarly across the entire range then focus can return to the $d = 5$ point estimate value. If designs are not robust for different error variance ratios, then trade-offs for different values will need to be considered explicitly.

This screening design was part of a sequence of experiments for understanding the response as a function of the factors, so any savings at this preliminary stage should enable additional resources to be available at later stages. The experimenter determined that the total number of observations, N, should be between 20 and 25, and the number of whole plots, $\#WP$, should range between 7 and $N - 1$, where the lower bound of 7 was selected because there are six whole plot terms in the model and it was desirable to have at least one degree of freedom for estimating the whole plot error variance. Because there was some flexibility about the overall size of the experiment, the relative advantages of a slightly smaller or larger experiment could be considered in the context of improving design performance within the cost and time constraints. In summary, the overall requirements for an appropriate design were as follows: (1) good D-efficiency; (2) robustness across the entire range of d values (i.e., $1 \leq d \leq 10$); (3) a manageably small $\#WP$; and (4) a small design size, N.

The remainder of the article is organized as follows: the next section provides some background about SPDs, optimal design construction and evaluation, as well as the Pareto front optimization approach for balancing multiple objectives. The following section presents the process for populating the Pareto front using an adapted point exchange algorithm and then selecting a design that best matches the priorities of the experimenter. The last section presents some general conclusions and discussion.

BACKGROUND

The Split-Plot Model and the D-Criterion

In split-plot experimentation, the following linear mixed model describes the variation in the $N \times 1$ response vector, $\mathbf{y}$,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon}. \quad [1]$$

Regarding notation, $\boldsymbol{X}$ is the $N \times p$ design matrix expanded to the model form for the p model parameters, $\boldsymbol{\beta}$, excluding the intercept, if applicable; $\boldsymbol{Z}$ is an $N \times \#WP$ classification matrix of ones and zeroes where the ijth entry is 1 if the ith observation ($i = 1, \ldots, N$) belongs to the jth whole plot ($j = 1, \ldots, \#WP$); $\boldsymbol{\delta}$ is a $\#WP \times 1$ vector of random effects where the elements are assumed independent and identically distributed (i.i.d.) $N(0, \sigma^2_{wp})$ with σ^2_{wp} denoting the variability among whole plots; $\boldsymbol{\varepsilon}$ is the $N \times 1$ vector of residual errors assumed to be i.i.d. $N(0, \sigma^2_{sp})$ with σ^2_{sp} denoting the variation among subplot units. It is also assumed that δ and ε are independent. Relating the model in [1] to the mixture-process experiment described earlier, $\mathbf{X} = [\mathbf{M} \mid \mathbf{P}]$, where $\mathbf{M}$ is of dimension $N \times 6$ and includes the linear whole plot mixture settings along with the three two-factor mixture-by-mixture interactions (these interactions are tested at the whole plot level). The matrix $\mathbf{P}$ is of dimension $N \times 6$ and contains the six two-factor mixture-by-process interaction terms. Note that the constraint of the mixture variables summing to one precludes including an intercept in $\mathbf{M}$ and the process main effects in $\mathbf{P}$ (due to the constraint that the sum of all two-factor mixture-by-process interactions for each of the process variable equals its corresponding main effect). For this example, we assume that there is no interaction among the process variables. This is a safe assumption for many applications. However, if there is prior knowledge or pilot data suggesting otherwise, then the appropriate process-by-process interaction terms can be added into the model.

The covariance matrix of the responses is

$$\begin{aligned} \mathrm{Var}(\mathbf{y}) = \boldsymbol{\Sigma} &= \sigma^2_{wp}\mathbf{Z}\mathbf{Z}' + \sigma^2_{sp}\mathbf{I}_N \\ &= \sigma^2_{sp}[d\mathbf{Z}\mathbf{Z}' + \mathbf{I}_N], \end{aligned} \quad [2]$$

where $\mathbf{I}_N$ is an $N \times N$ identity matrix and $d = \sigma^2_{wp}/\sigma^2_{sp}$. For simplicity of presentation, we sort the observations by whole plots, to obtain $\mathbf{Z} = diag\{\mathbf{1}_{n_1}, \cdots, \mathbf{1}_{n_{\#WP}}\}$, where $\mathbf{1}_{n_j}$ is an $n_j \times 1$ vector of one's and n_j is the size of the jth whole plot. The diagonal form of $\mathbf{Z}$ induces a block-diagonal structure for the covariance matrix, Σ, with $\Sigma = \mathrm{diag}\{\Sigma_1, \ldots, \Sigma_{\#WP}\}$ where each $n_j \times n_j$ matrix Σ_j is given by

$$\Sigma_j = \begin{bmatrix} \sigma^2_{sp} + \sigma^2_{wp} & \cdots & \sigma^2_{wp} \\ \vdots & \ddots & \vdots \\ \sigma^2_{wp} & \cdots & \sigma^2_{sp} + \sigma^2_{wp} \end{bmatrix} = \sigma^2_{sp} \begin{bmatrix} 1+d & \cdots & d \\ \vdots & \ddots & \vdots \\ d & \cdots & 1+d \end{bmatrix},$$

and Σ_j denotes the covariance matrix of responses in the jth whole plot. Note that the variance of an individual observation is the sum of the subplot and whole plot error variances, $\sigma^2_{sp} + \sigma^2_{wp}$. A popular method for estimating the variance components is restricted maximum likelihood (see Searle et al. 1992, chapter 6, for details on restricted maximum likelihood estimation).

The generalized least squares estimate of the fixed effects parameters $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\boldsymbol{y}$, with the covariance matrix of the estimated model coefficients as

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}. \quad [3]$$

From [3] we note that the information matrix associated with the SPD is $(\mathbf{X}'\Sigma^{-1}\mathbf{X})$. The D-criterion is defined as the determinant of the information matrix and is given by

$$\begin{aligned} |\mathrm{I}(\xi)| &= |\mathbf{X}'\Sigma^{-1}\mathbf{X}| = \left|\left(\sigma^2_{wp} + \sigma^2_{sp}\right)(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})\right| \\ &= (1+d)|\mathbf{X}'[d\mathbf{Z}\mathbf{Z}' + \mathbf{I}_N]^{-1}\mathbf{X}|. \end{aligned} \quad [4]$$

In [4], $R = \frac{1}{\sigma^2_{wp} + \sigma^2_{sp}}\Sigma$ denotes the observational correlation matrix. A design's D-efficiency is then defined as $|\mathbf{I}(\xi)|^{1/p} = (1+d)^{1/p}|\mathbf{X}'[d\mathbf{Z}\mathbf{Z}' + \mathbf{I}_N]^{-1}\mathbf{X}|^{1/p}$. Since the determinant in [4] is inversely proportional to the square of the volume of the confidence region for the estimated regression coefficients, the D-criterion is a single number summary reflecting the overall precision of the estimation of $\boldsymbol{\beta}$ with large values corresponding to lower variances and hence more precise estimation. A key feature of design optimization when using the D-criterion for the split-plot case is the fact that the D value depends on the variance components ratio, d (Goos 2002), the structure of $\mathbf{X}$, which is a function of whole-plot levels and the arrangement of subplot levels within whole plots, and the

dimensionality of each of the Σ_j (determined by the number of subplots in each whole plot).

Optimal Design Construction and Evaluation

The construction and evaluation of optimal SPDs have received considerable attention in the literature. Huang et al. (1998) and Bingham and Sitter (1999, 2001) use *minimum aberration* criteria for determining optimal two-level fractional factorial screening split-plot experiments. Minimum aberration designs achieve the least amount of departure from an orthogonal design for a given set of restrictions (e.g., fixed number of whole plot factors, fixed number of subplot factors, whole plot size, etc.). Goos and Vandebroek (2001, 2003, 2004) proposed point exchange algorithms for determining D-optimal split-plot designs. Liang et al. (2007) developed cost-penalized expressions for D-, G-, and V-optimality criteria. Jones and Goos (2007) demonstrated a coordinate exchange algorithm with improved efficiency for selecting D-optimal split-plot designs over point exchange algorithms. In Goos and Vandebroek (2003) and Jones and Goos (2007), whole plot sizes are constrained to be equal but this constraint is relaxed in Goos and Vandebroek (2001, 2004). Update formulas for improving the efficiency of point exchange algorithms involving 4+ factors with balanced or unbalanced whole plot sizes were developed by Arnouts and Goos (2010). Smucker et al. (2011) developed a coordinate exchange algorithm for constructing high D-efficiency split-plot designs robust to model misspecification within a specified family of models. Trinca and Gilmour (2001) proposed designs for restricted randomization cases built by strata. Parker et al. (2006) developed a class of split-plot designs in which the ordinary least squares estimates of model parameters are equivalent to the generalized least squares estimates. Anbari and Lucas (1994) considered the G-criterion for comparing competing split-plot designs. Liang, Anderson-Cook, and Robinson (2006) and Liang, Anderson-Cook, Robinson, and Myers (2006) considered graphical techniques for assessing competing split-plot designs.

Though many criteria have been developed for determining an optimal SPD, rarely does a universally "best" design exist that outperforms all competitors for all criteria of interest. A good design that produces reliable results under a wide variety of user-specified objectives is highly desirable. For a summary of important design characteristics, see Box and Draper (1975) and Myers et al. (2009, p. 282). D-efficiency has been one of the most commonly used criteria for design selection, which will be considered as the main characteristic of design performance for our particular case study. In general, there are often other qualities of interest that should also be considered (see Robinson and Anderson-Cook 2011).

A unique challenge for selecting an optimal SPD is that many of the design characteristics associated with SPDs (including D-optimality) depend on the size of the generally unknown whole plot–to-subplot variance ratio, $d = \sigma^2_{wp}/\sigma^2_{sp}$. Software for constructing D-optimal SPDs requires the user to specify a model, a value for d, and the total number of observations, N. Hence, the user is often faced with the task of specifying a *best guess* of this ratio. However, it is often difficult to precisely specify a particular value of d, especially when data from a pilot study on the process are sparse or nonexistent. It is often more realistic to ask the experimenter to specify a range of possible values for d and this was the approach taken here because the experimenter was not confident in completely relying on the pilot data's suggestion that $d = 5$. Because the choice of a D-optimal design may be dependent on this ratio, it is important to consider this uncertainty during design construction. Therefore, obtaining a D-optimal design from a statistical software package for a particular value for d does not necessarily guarantee good performance, across all values of d. With multiple objectives considered simultaneously, selection of an optimal design should be based on carefully balancing competing objectives by examining the trade-offs between the criteria of interest.

The Pareto Approach

Pareto multiple objective optimization has been used broadly in many disciplines in the past decade. However, only recently was the Pareto method introduced to the design of experiments paradigm (Lu et al. 2011, 2012; Park 2009). The desirability function (DF) approach of Derringer and Suich (1980) is currently a common approach for multiple criterion design optimization. The DF method requires the user to determine the relative importance of individual criteria of interest a priori and subsequently combine all criteria values into a single index based on these priorities. The

DF approach does not directly consider the trade-offs between criteria but rather identifies a single best design by optimizing the summary index. In addition, the DF solution can depend heavily upon the user-specified weights, scaling scheme (to convert criteria value onto the DF scale), and the DF metric form (to integrate multiple criteria into a single summary). Note that all of these choices are subjective and require expert knowledge and/or information on the criteria value ranges prior to the search for a design. In practice, it is not unusual for experts to disagree on these choices, thus making obtaining a single set of weights for the DF approach problematic.

In this article, we describe the Pareto front approach by Lu and Anderson-Cook (2014) within the split-plot setting. The approach assembles a suite of candidate designs using Pareto optimality (Gronwald et al. 2008; Kasprzak and Lewis 2001). The approach consists of two stages: (1) An objective stage focuses on developing the Pareto frontier (or front) consisting of contending designs where poor designs have been eliminated and no remaining designs are strictly better than members of the set for all criteria; and (2) a subjective stage that compares the selected designs from the first stage using weighting combinations that reflect experimenter priorities. This second stage is called the Pareto decision analysis and involves selecting a subset of designs from the Pareto front based on examining the trade-offs of the designs, design performance, and robustness to different weighting schemes.

One major advantage of the approach is that the computationally intensive optimization search is completed only once at the start of the process. All subsequent analyses are computationally inexpensive using the results of the single, initial search but allow the user to have great flexibility when exploring different weighting choices. The Pareto approach for design selection was introduced for the completely randomized setting in Lu et al. (2011). An adapted point exchange algorithm for Pareto optimality within the split-plot setting with fixed and flexible numbers of whole plots was developed by Lu and Anderson-Cook (2014).

The identification of the Pareto front, through the Pareto optimization search process, provides an objective set of non-inferior choices before introducing any subjective aspects. It allows the practitioners to see all of the contending choices and understand the range of values and trade-offs across the criteria. When there is uncertainty associated with the subjective choices, the Pareto approach requires little extra effort to conduct a sensitivity analysis for evaluating the impacts of subjective inputs once the Pareto front is identified. It allows the experimenter to make and evaluate sensible choices on the weighting and scaling schemes as well as the DF form. It is important to note that when there are multiple weighting, scaling, and DF form combinations of interest, the DF approach is computationally expensive to evaluate because every new setup (combination of choices) requires a new search.

The key to successfully using the Pareto approach is to efficiently populate the complete Pareto front. For this study, we utilized the Pareto aggregating point exchange for split-plot designs (PAPESPD) algorithm of Lu and Anderson-Cook (2014). For each design size and a fixed #*WP*, parallel searches for many random starting designs are used to strategically guide the searches in diverse directions. The use of multiple weight combinations improves the coverage of the entire front and prevents the search from getting stuck at local optima (see Lu and Anderson-Cook [2012] for more details). Because there may be many isomorphic designs identified, if two designs are found with exactly the same values for all of the criteria of interest, only the first of these is retained. Though the approach utilized here relies on a fixed #*WP* for each design size, the PAPESPD can also be implemented to handle a flexible #*WP* (Lu and Anderson-Cook 2014).

The second stage, the Pareto decision analysis, involves ranking the designs on the Pareto front in terms of the proximity of each design's criterion vector to the criterion coordinates associated with the Utopia point. The Utopia point in the criterion space has coordinates that are the best possible found for each criterion from all the searches. Although the Utopia point usually only exists in a theoretical sense, it can be used as an "ideal" target to calibrate closeness based on a chosen metric. A common distance metric for the Utopia point approach is the L_1-norm formulated as

$$\underset{\xi \in \Omega^*}{\text{Min}} \sum_{j=1}^{C} |w_j \left(f_j(\xi) - f_j^0 \right)|, \qquad [5]$$

which matches the use of an additive desirability function. In the above formula, Ω^* denotes the set of designs on the Pareto front, ξ is a design on the front, $f_j(\xi)$ is the scaled objective or desirability function for

the jth criterion corresponding to design ξ, and f_j^0 denotes the Utopia point value for the jth criterion. The weights w_j in $j \in \{1, 2, \ldots, C\}$ reflect the user-specified preferences of the criteria. If the multiplicative DF is of interest, then it can be converted to an additive form with a logarithm transformation. This matches the use of the L_1-norm on the log scale of the Utopia point approach. The choice of which DF form to use depends on whether the experimenter wants to severely penalize poor performance of any criterion (multiplicative DF) or whether stellar performance in one criterion is thought to be a reasonable trade-off for very poor performance on the other (additive DF).

Designs that are best for at least one set of weights can be further screened by (1) studying the robustness of the designs to ranges of weighting combinations; (2) examining the individual attributes of each design along with the trade-offs between criteria among the designs; and (3) studying the performance of each design relative to the best possible design for each of the weighting combinations specified by the experimenter as most desirable. The final decision on design choice is based on the priorities of the study combined with the relative performance of the most promising designs. The Pareto approach offers the user a flexible method for examining important trade-offs between the multiple design objectives.

FINDING BEST DESIGNS FOR THE FILM EXAMPLE

We now describe the Pareto front approach as it applies to the film making design selection problem. We defined the first two criteria as the D-criterion values associated with the variance ratios $d = 1$ and $d = 10$ and we denote these as D(1) and D(10) respectively. In terms of the notation presented previously, $f_1(\xi) = \mathrm{D}(1)$ and $f_2(\xi) = \mathrm{D}(10)$. The motivation for considering both D(1) and D(10) as separate criteria was based on results from Goos (2002) and Lu and Anderson-Cook (2014), who demonstrated that the choice of a best design is dependent on the value of d selected for many split-plot cases involving continuous factors.

In addition we considered the number of whole plots, $\#WP$, and the total size of the experiment, N, as other criteria of interest, and we denote these as $f_3(\xi)$ and $f_4(\xi)$, respectively. The number of whole plots and the total design size were kept as separate objectives, because changing the $\#WP$ is primarily tied to the time required to run the experiment, while N is mostly associated with the overall cost of the experiment. This differs from approaches suggested by Bisgaard (2000) and Liang et al. (2007), who combine the overall cost of the experiment (including time and financial cost) into a single measure. Keeping the criteria separate enables one to see trade-offs between these criteria from different options on the Pareto front and to disentangle the individual contributions from cost and other design characteristics and evaluate their trade-offs and impact on decision making.

Objective Phase—Finding the Noninferior Designs on the Pareto Front

For the study presented here, each possible competing design, ξ, is associated with a 4×1 criterion vector, $\mathbf{F}(\xi) = (f_1(\xi), f_2(\xi), f_3(\xi), f_4(\xi))'$, where $f_1(\xi) = \mathrm{D}(1)$, $f_2(\xi) = \mathrm{D}(10)$, $f_3(\xi) = \#WP$, and $f_4(\xi) = N$. Recall that the first step in the Pareto approach requires the user to find the set of all non-inferior designs according to the definition of Pareto optimality. This stage is the computationally expensive stage and to accomplish this in a computationally efficient manner, we utilized the PAPESPD of Lu and Anderson-Cook (2014) separately for each combination of design size, N (where N ranged from 20 to 25) and $\#WP$ (where $\#WP$ ranged between 7 and $N - 1$) using the criterion vector $\mathrm{F}(\xi) = (\mathrm{D}(1), \mathrm{D}(10))'$. Multiple starts were utilized for each combination of N and $\#WP$ and the 28 design locations in Figure 1 were considered as the candidate set for the point exchange. Once these individual fronts are generated, the results were combined to give the overall four-criteria Pareto front. Each of the two-criteria Pareto fronts was generated using six parallel searches for each random start where the six sets of weights were chosen to consider criterion weight values that were nonnegative multiples of 0.2 with the two weights summing to 1. Each two-criteria front takes approximately 90 to 190 min to evaluate the different combinations of N and $\#WP$ with 300 random starts. The decision regarding when to stop the search algorithm (running more random starts will not improve the Pareto front) was based on monitoring the Pareto front growth using the scaled hypervolume (the area

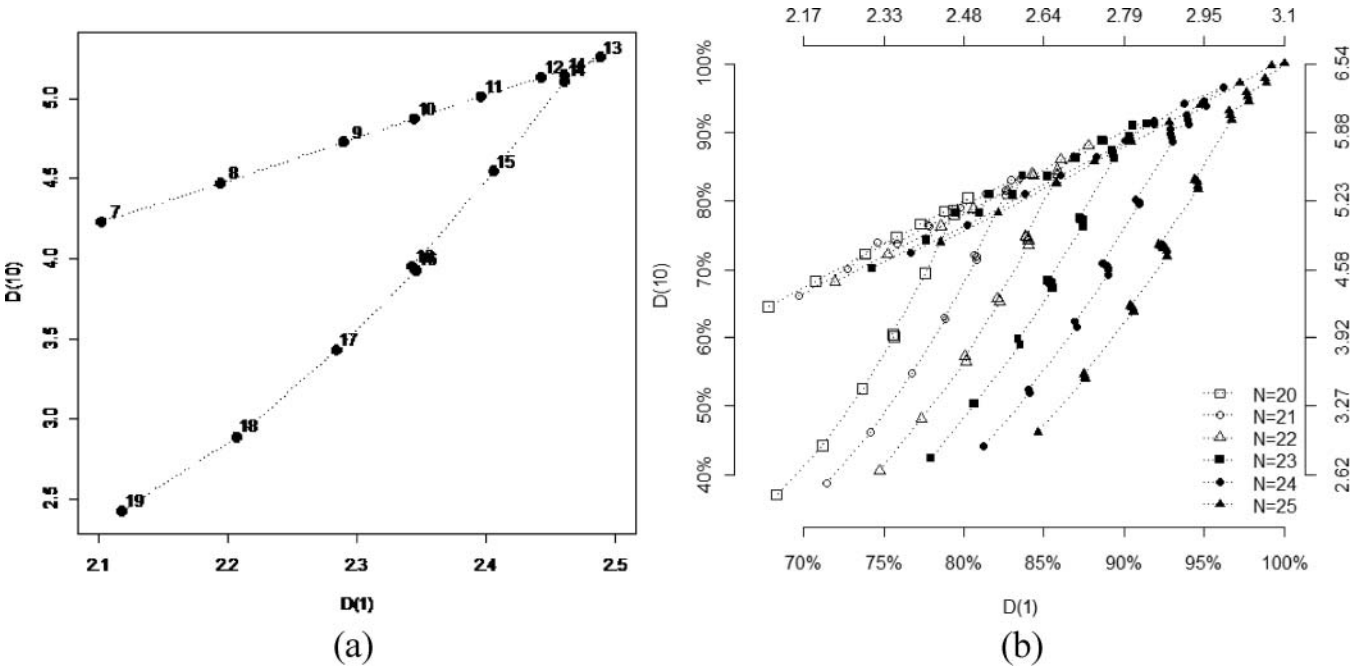

FIGURE 2 **Pattern of best designs based on D(1) and D(10) criteria for the mixture-process example for (a) $N = 20$ and (b) $N = 20$–25 and different *#WPs*.**

under the Pareto front for a two-criteria case; see Lu and Anderson-Cook, 2013).

Figure 2a shows a scatterplot of the coordinates of the criterion vector $\mathbf{F}(\xi) = (D(1), D(10))'$ for each of the Pareto front designs associated with every combination of $N = 20$ and $\#WP = 7$–19. From the figure, note that the optimal $\#WP$ for maximizing D-efficiency when $N = 20$ is 13 for both $d = 1$ and $d = 10$. For $\#WP$ less than 13, there is a trade-off between reducing the time for the experiment ($\#WP$) and the D-efficiency. From $\#WP = 7$ to 13, using more whole plots resulted in improved estimation of the model parameters but at the price of longer experimental time. For $\#WP > 13$, there is no improvement in the D-efficiency and the time required to run the experiment continues to increase. Hence, these designs were not sensible to consider further.

Figure 2b provides a scatterplot of the coordinates of the criterion vector $\mathbf{F}(\xi) = (D(1), D(10))'$ for each of the Pareto front designs when considering all of the combinations of N and $\#WP$, with separate lines for each value of N. The range of D-efficiencies shown on the left and lower axes differs considerably, with far more spread in the range of efficiencies for D(10) [left axis], which ranges between 37% (when $\#WP$ is nearly equal to N) and 100%. On the other hand, the range of D(1) [bottom axis] only varies from 67.8 to 100%. This indicates with an appropriately chosen $\#WP$, the SPD can achieve substantial improvement in D-efficiency relative to the completely randomized design (CRD) (note when $\#WP = N$, the SPD is just a CRD) and the improvement is more dramatic for bigger d values. We note that the shape of the lines is quite consistent across all design sizes with nearly constant rates of improvement in D-efficiency as we increase either N or $\#WP$. In addition, for each line there is a single best $\#WP$ that is optimal for both D(1) and D(10). This surprising result differs from what is typical in many split-plot experiments where different *#WPs* are best for different values of d. For example, in the Appendix, we show an example of another split-plot experiment with five continuous non-mixture factors (three whole plots and two subplot factors) for $N = 20$ in Figure A1(a). In this case, the best $\#WP$ differs depending on whether we consider D(1) or D(10). For D(1), the best *#WPs* is 10 and 16, for $N = 20$ (a) and 25 (b), respectively. In contrast, the corresponding *#WPs* that maximize D(10) are 9 and 12, respectively. Note that the robustness of the optimal mixture-process example to choices of d could not have been anticipated a priori because this robustness is often not present in other SPDs. The shape of the combined Pareto fronts for a given design size for this class of mixture-process experiments holds across a broad range of experimental setups that the authors have explored.

Figure 3 shows parallel plots of the combined four-criteria Pareto front with 60 designs for $\mathbf{F}(\xi) = (D(1), D(10), \#WP, N)'$ with D(1) and D(10) shown adjacent to each other and with the same scale based on the worst efficiency (of 64.7% from D(10)) from either criteria mapping to zero on the desirability scale. The fact that the two estimation-based criteria perform so similarly yields the nearly parallel lines connecting these two criteria. This robustness of the D-criterion to the range of specified variance ratios is unique to the class of models considered for our problem. For example, Figure A2(b) in the Appendix shows the continuous case with five factors, where the lines connecting D(1) and D(10) are substantially less parallel, indicating more trade-offs between these criteria and a stronger dependence on the user correctly specifying the variance ratio value. In the non-mixture example case, it is necessary to more proactively manage how the uncertainty from the variance ratio influences other characteristics of the designs and the experimenter's decision.

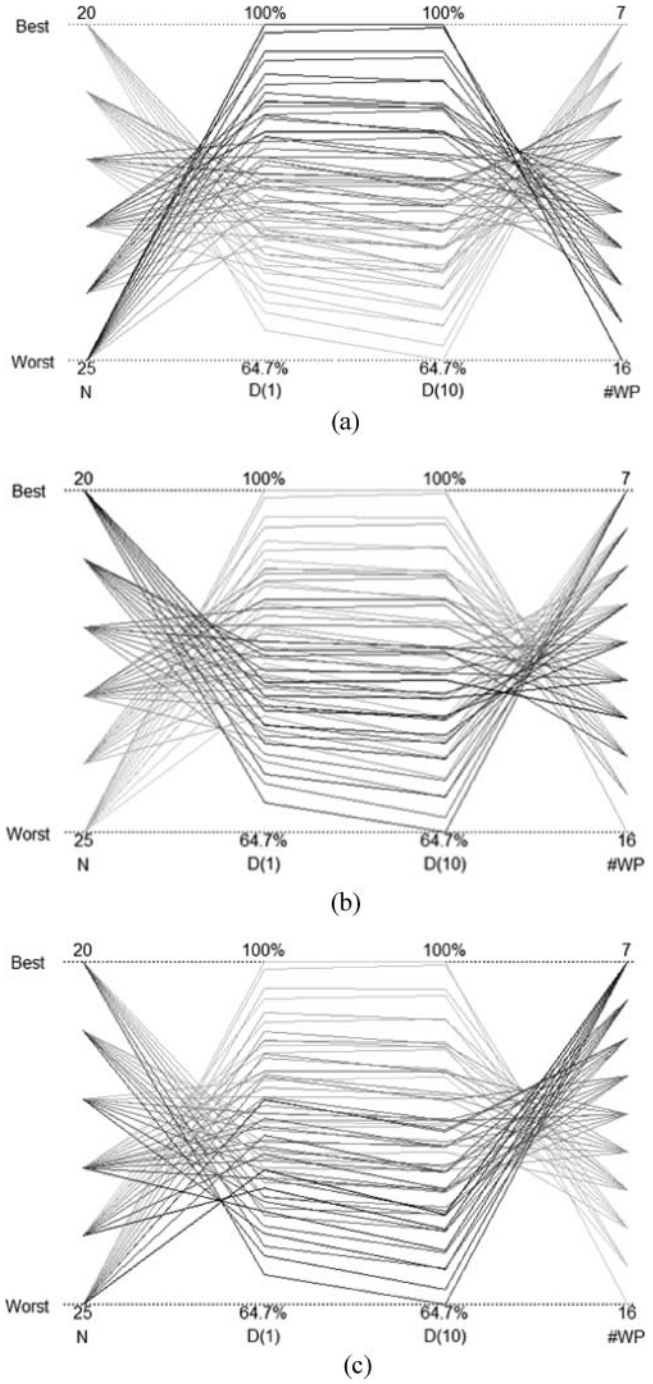

FIGURE 3 **Parallel plots based on prioritizing (a) D(1) and D(10), (b) *N*, and (c) *#WP* with darker lines for better performance for the prioritizing criterion. These plots illustrate the sharp trade-offs required between the D-efficiency criteria and the cost-based measures.**

To consider the effects of valuing the criteria differently, Figure 3a highlights the designs with better performance for the D-efficiencies with darker lines. Note how the best designs here correspond to the larger design sizes ($N = 25$) and *#WPs* ranging between 15 and 16. If minimizing the financial cost of the experiment is of top priority, then Figure 3b highlights the range of performance attainable for D(1) and D(10) for $N = 20$. Finally, Figure 3c shows what range of design performance is possible if the experimenter prioritizes minimizing the time to run the experiment with *#WP* set to the minimum.

Though the impact of a misspecified variance ratio can be problematic in many SPD settings, in our example we found the optimal design to be robust to the choice of *d*. Because the D-efficiency was found to be similar across the entire range of considered variance ratios (i.e., $(1 \leq d \leq 10)$), we changed our estimation based criteria to D(5), which corresponds to the initial estimate of *d* from the pilot study. Hence, the optimization problem was adjusted to find the three-objective Pareto front using $\mathbf{F}(\xi) = (D(5), \#WP, N)'$ with the Utopia point values of 4.93, 7, and 20, respectively. It should be noted that in general it is hard to predict the impact of a misspecified variance ratio and hence the robustness to the variance ratio should be carefully evaluated whenever one wishes to find an optimal SPD.

Subjective Phase—Selecting a Best Design to Match Experimenter Priorities

We now describe the subjective selection of a best design from those identified on the Pareto front, by incorporating the experimenter's requirements and

priorities for the study. Recall that the choice of the multiplicative or additive DF form specifies the amount of penalty for poor performance by one or more of the criteria. Here the experimenter was willing to consider poor performance of any of the criteria if it was accompanied by exceptional performance of another, and hence the additive DF form and the L_1-norm of Eq. [5] were utilized.

Because often users are more comfortable defining ranges of weights for each criterion, we considered a fine mesh (200 gradations for each w_j between 0 and 1) of possible combinations of the w_j's and for each combination, the 51 designs on the three-criteria Pareto front were ranked according to the evaluated expression in [5]. Because we require $\sum_{j=1}^{3} w_j = 1$, the best designs for all possible weight combinations can be displayed with the mixture plot shown in Figure 4. This utilizes the simplex coordinate system (see Cornell 2002, p. 24) for mixtures of ingredients. The three vertices correspond to optimizing based on a single criterion, the three edges consider each pair of the criteria, and the interior of the simplex has non-zero weights for all three criteria.

Of the 51 designs on the three-criteria Pareto front, 15 designs are optimal for at least one combination of the weights using the L_1-norm. The weight combinations for which a design is optimal are shown in grayscale in Figure 4. Assuming that the user wishes to select a design that is robust to a range of weight combinations, we highlight nine designs that are best for at least 1% of the total area of the simplex and their numeric labels are assigned from largest to smallest value of D(5). The D-efficiencies and their corresponding area in the mixture plot for the nine designs are shown in Table 1. Note how the ranking of the best

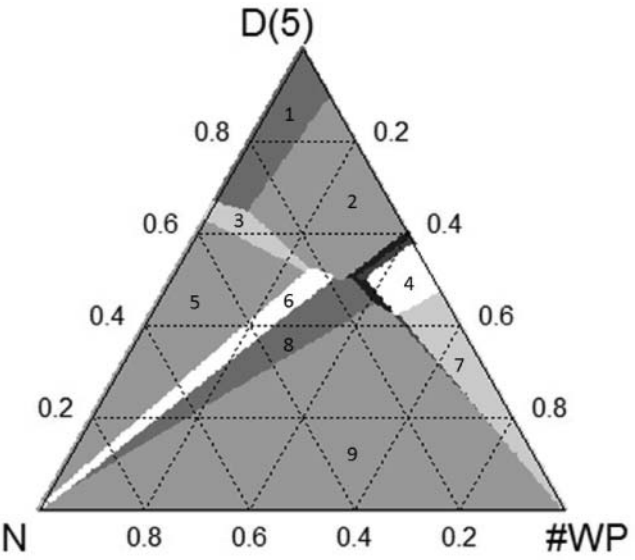

FIGURE 4 **Mixture plot based on three criteria: D(5), *N*, *#WP*.**

designs for D(5) is consistent with the rankings for both D(1) and D(10), as well as that the D(5) D-efficiency values are bounded by those for D(1) and D(10).

Figure 4 shows that Design 1, with $N = 25$ and $\#WP = 16$, is D-optimal for $d = 5$. However, we also observe that by sacrificing 0.5% in D-efficiency, utilizing Design 2 (with $N = 25$ and $\#WP = 15$) results in a 6.7% savings in terms of experimental time. Design 2 is also optimal for 12.1% of the possible weighting combinations and thus is more robust to the user's choice of weighting combinations. Design 5 ($N = 20$, $\#WP = 13$) and Design 6 ($N = 20$, $\#WP = 10$) also present nice alternatives if one is willing to opt for greater cost-efficiency through reduced experimental time or reduced financial cost. More specifically, Designs 5 and 6 have lower D-efficiency with values of 80.4 and 74.7%, respectively, but they also result in fewer total observations and fewer *#WP*. If low costs (for both

TABLE 1 **Designs Selected by the Adapted Utopia Point Approach with Criteria Values and Area in the Simplex (at least 1%) Using the L_1-Norm (equivalent to the additive DF)**

Design	*N*	*#WP*	Rel. D(5) (%)	Area (%)	Rel. D(1) (%)	Rel. D(10) (%)
1	25	16	100	5.7	100	100
2	25	15	99.5	12.1	99.2	99.7
3	24	15	96.5	1.4	96.2	96.6
4	25	9	83.0	2.1	85.8	82.7
5	20	13	80.4	15.8	80.2	80.4
6	20	10	74.7	2.9	75.8	74.7
7	25	7	74.5	5.0	78.5	74.0
8	20	9	72.4	7.8	73.8	72.3
9	20	7	65.0	44.6	67.8	64.7

time and financial cost) are the top priorities, then Design 9 is preferred for a large portion of the simplex. One advantage of the Pareto approach is that the range of criteria values on the front provides some calibration of what values are sensible to consider. For example, by constructing the front, we learned that there is no advantage to running a design with more than 16 *WPs* and that we can achieve good efficiency with several less expensive alternatives.

The trade-off plot (Lu et al. 2011) in Figure 5 shows the 15 designs that are best for at least one combination of weights, with the nine designs from Table 1 denoted with bold black symbols. The designs are sorted from left to right for increasing values of D(5). An interesting aspect of the L_1-norm is that it tolerates poor performance of one criterion if there is very good performance from another. This results in designs with extreme values of N being selected. If the L_1-norm on the log scale (equivalent to the multiplicative DF) had been selected by the experimenter to combine the different criteria, then designs with intermediate values of N would be selected (results available from the authors by request). This emphasizes the subjective nature of the second phase to select a design using the Pareto front approach with the flexibility to accommodate different priorities. In general, the experimenter should think carefully about which form and weightings make the most sense for the study goals.

Because the overall goal was to determine a single design for implementation, it was advantageous to consider how much of a compromise different designs required from the best possible choice for given weighting combinations. To help evaluate this type of a trade-off, Lu and Anderson-Cook (2012) developed the synthesized efficiency plot. The synthesized efficiency for a particular design is defined as

$$SE(\xi, \boldsymbol{w}) = \frac{DF(\xi, \boldsymbol{w})}{max_{\xi}(DF(\xi, \boldsymbol{w}))},$$

where $DF(\xi, \boldsymbol{w})$ is the additive DF value for a design ξ and a specific weight combination $\boldsymbol{w}$. The synthesized efficiency plot then displays $SE(\xi, \boldsymbol{w})$ for all weight combinations in the weighting simplex. These plots are displayed in Figures 6a–6d for Designs 2, 5, 6, and 9, respectively. In the plots, white and light gray correspond to good efficiency, and darker shades of gray show poor efficiency relative to the best possible design at a given weight combination. The plots highlight the strong trade-offs between good D-efficiency and the cost (time and financial) of the design. Designs 2 and 9 perform very well for either D-efficiency (Design 2) or the cost (Design 9) and hence there is a region with high synthesized efficiency (white or light gray), which is close to the regions where the design is best. However, there are also regions with quite poor performance (dark gray

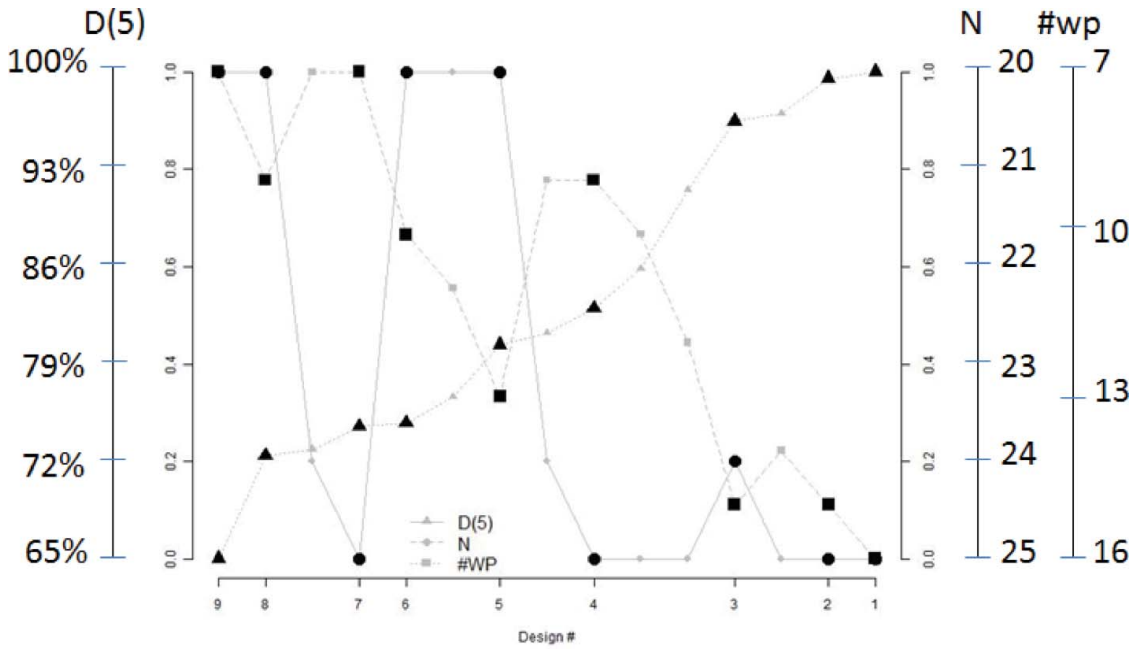

FIGURE 5 Trade-off plot for designs selected by the Utopia point approach based on the L_1-norm sorted by relative D(5)-efficiency.

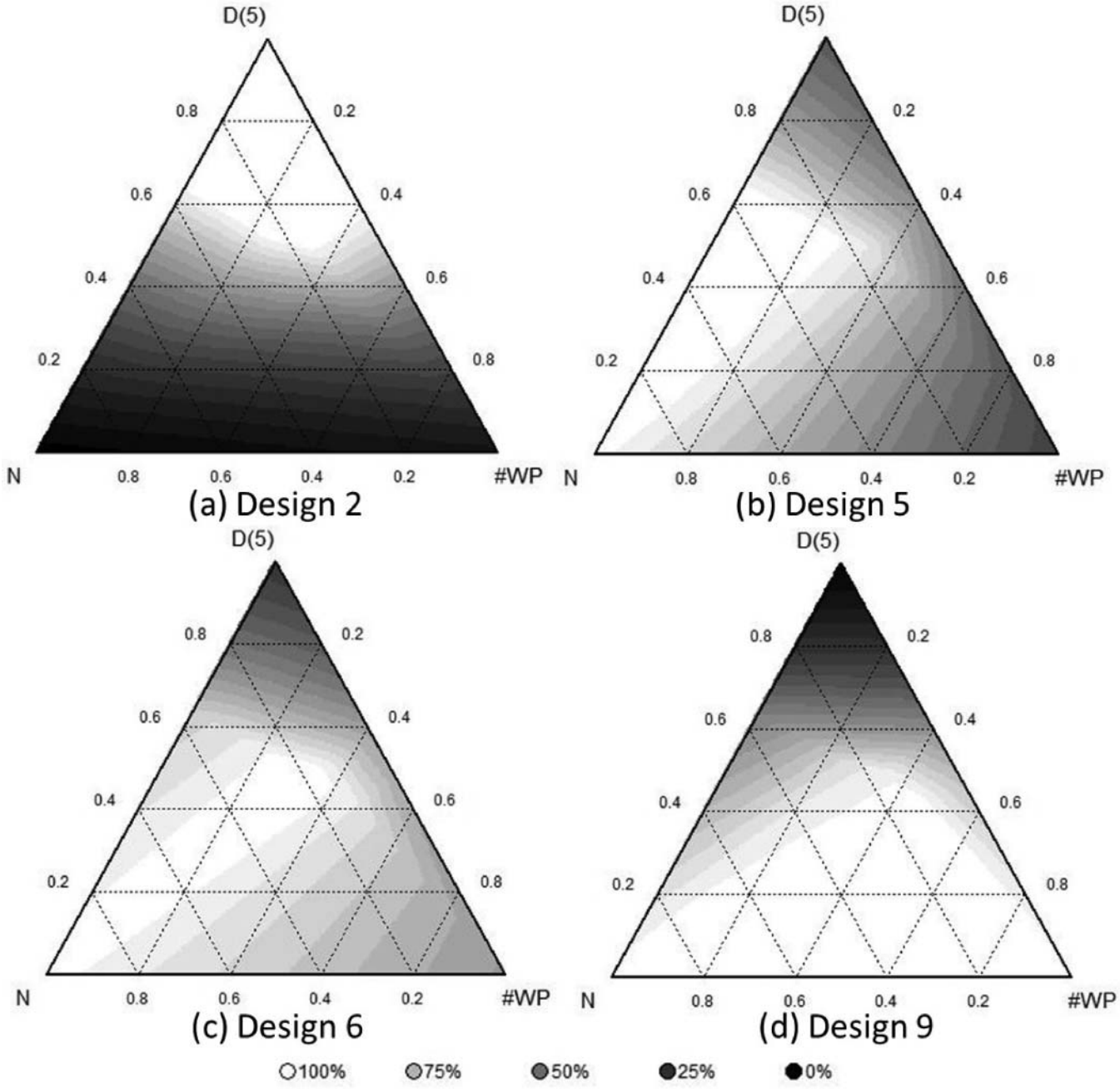

FIGURE 6 **Synthesized efficiency plot for designs (a) 2, (b) 5, (c) 6, and (d) 9. The white–gray–black scale represents high to low synthesized efficiency.**

or black). Designs 5 and 6 have more balanced performance with a worst-case synthesized efficiency no worse than 27% for all possible weight combinations. Specially, Design 6 has 78% of the area with at least 75% synthesized efficiency, which means that this design performs quite well for the majority of the possible weights. However, no single design is able to simultaneously perform well for all three objectives, which emphasizes the need for experimenters to think carefully about how to prioritize the study goals.

Finally, Figures 7a–7d show the geometric representations for Designs 2, 5, 6, and 9, respectively. This figure allows the experimenter to see which combinations of the three whole plot/mixture factors (X_1, X_2, X_3) and two subplot/process factors (P_1, P_2) should be run. The numbers within the candidate locations are labels for the whole plots, and to run the experiment,

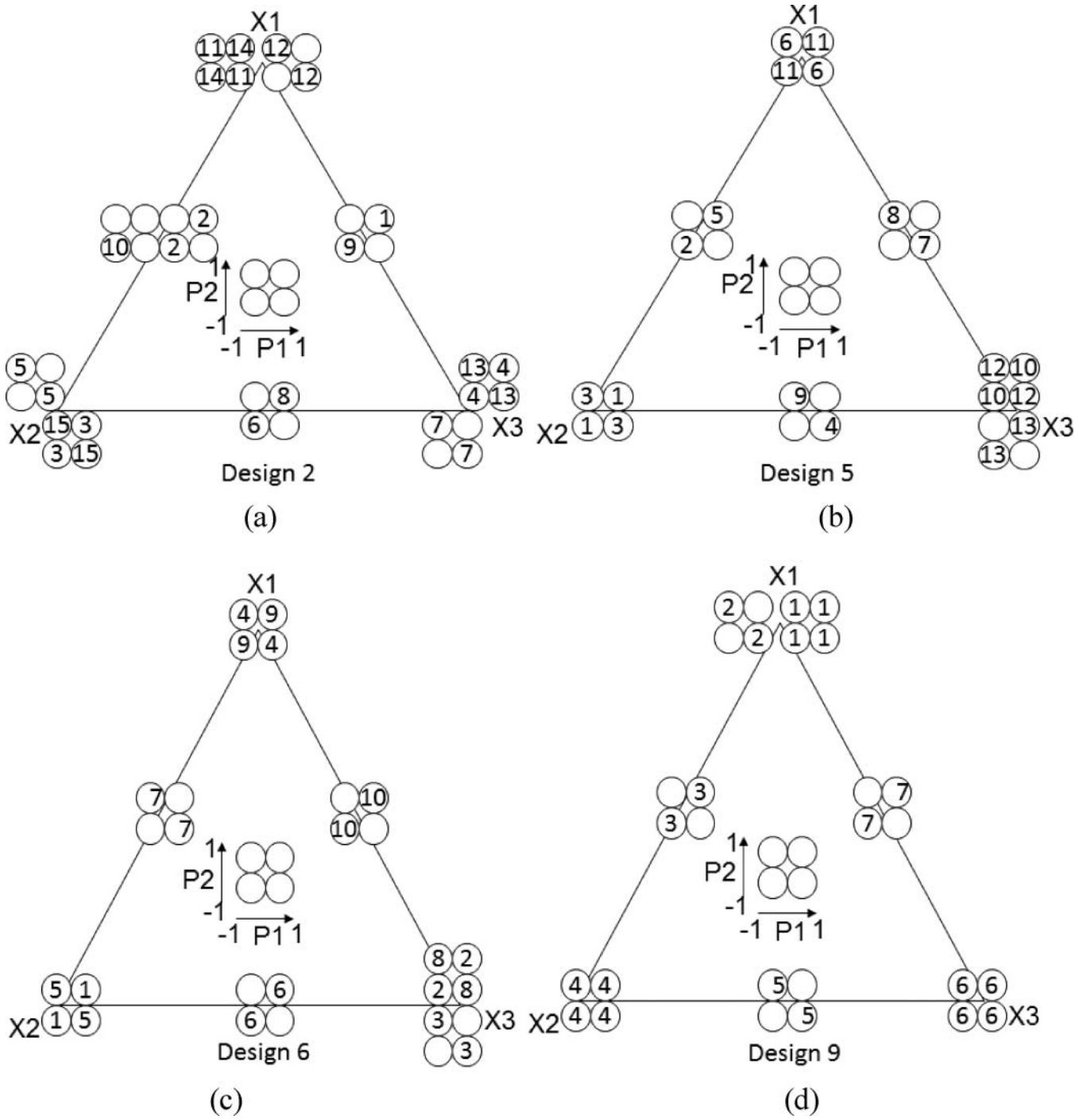

FIGURE 7 Geometry plot for designs (a) 2, (b) 5, (c) 6, and (d) 9.

two separate randomizations should be performed: one to determine the order in which the whole plots are run, and the second randomization determines the order of subplot combinations within each whole plot. When the design involved replication of factor combinations, the candidate locations at a given whole plot location are shown with multiple groups. None of the designs shown here have any observations located in the centroid of the mixture factor candidate space. This matches the nature of the D-criterion in the sense that extreme points in the design space are favored over points within the interior of the design region. Design 2 ($N = 25$ and $\#WP = 15$) has three whole plots, each with two observations at the corners of the mixture

candidate space and the remaining seven observations are spread between six whole plots on the edges of the mixture space. Designs 5, 6, and 9 (all with $N = 20$ and $\#WP = 13$, 10, and 7, respectively) have 14 observations at the vertices of the mixture candidate space and 6 observations on the edges with balance between the number of observations with the high and low levels of the subplot factors. It is worth noting that Design 6 exhibits the best symmetry among the four designs and it is balanced (i.e., each whole plot is of size 2).

Based on the graphical summaries included in the Pareto decision analysis, the experimenter was presented with the necessary information for understanding the possible alternatives. Design 2 clearly has the best estimation precision, but choosing this design comes at the cost of an expensive and time-consuming design because it requires the maximum design size ($N = 25$) and greatest number of whole plots ($\#WP = 15$). At the other extreme, Design 9 is cheaper and less time-consuming to run, but choosing this design results in a sacrifice of the D-efficiency. For this example, the experimenter felt that both of these extremes were undesirable and thus focused upon Designs 5 and 6. These two designs are relatively similar and both have the same design size ($N = 20$). The primary difference in these two designs is in the number of whole plots to be utilized and the final decision was based upon whether the experimenter believed that the time savings represented by three fewer whole plots is worth the approximately 5.7% reduction in D-efficiency for $d = 5$ (5.7% for $d = 10$ and 4.4 for $d = 1$). In this case, the experimenter felt that the additional whole plots of Design 5 did not offer sufficient improvement in performance to warrant the additional cost and hence Design 6 was selected. We emphasize that there is no single correct answer regarding how to value these trade-offs, but a major benefit of the Pareto front approach is that it clearly highlights the sacrifices that one is making with one criterion to improve another. Another experimenter could quite rationally have selected any of the designs in Table 1 as a best match for the needs of the study design.

The choice of an optimal design should involve a process of understanding the available choices, observing trade-offs between the criteria of interest for promising designs, and a subjective determination of a preferred range of weights associated with each criterion. The goal of the Pareto front approach is to provide the set of sensible alternatives in the objective phase and then to provide numerical and graphical summaries that equip the decision maker to choose what is best for the study goals.

SUMMARY AND CONCLUSIONS

Computer-generated, optimum SPDs have traditionally been based on a single, user-defined design optimality criterion. Parker et al. (2008) and others pointed out that the choice of an optimal SPD should simultaneously involve the use of multiple design criteria. Though other research considers multiple criteria for determining an optimal SPD (Goos 2002; Liang et al. 2007; Parker et al. 2008; Smucker et al. 2011), only the Pareto front approach (Lu and Anderson-Cook 2014) can be utilized for any flexible set of user-supplied design criteria. In this article, the Pareto front approach was illustrated with a mixture-process example involving five factors (three mixture whole plot and two process subplot factors). The development of the Pareto front was conditioned on a candidate set of design points and a user-specified model.

In this case, the practitioner sought to find an optimal SPD that balanced good statistical estimation properties along with practical concerns such as cost and time constraints. Financial cost was measured by design size, N, whereas time constraints were quantified by $\#WP$. In terms of statistical properties, the practitioner was interested in estimating the model coefficients as precisely as possible and hence the D-criterion was of interest, which is dependent on the unknown whole plot–to-subplot variance ratio, $d = \sigma^2_{wp}/\sigma^2_{sp}$. In many applications, precise specification a single value of d is difficult and the experimenter is more likely to be able to provide a range of possible values. Here the experimenter sought to protect against possible misspecification of d. Specifically, the experimenter wished to find a highly D-efficient design that was robust to the entire range of anticipated variance component ratios (i.e., $1 \leq d \leq 10$). After initial exploration of the four criteria Pareto front using the D(1) and D(10) efficiencies, N, and WP, it was discovered that the D-optimal design for a fixed N and $\#WP$ was the same for both $d = 1$ and $d = 10$. This unique phenomenon for mixture-process SPDs, however, is different from many other continuous non-mixture scenarios where the best design is highly dependent on the chosen value of d. Given the robustness across D(1) and D(10) efficiencies, the four-criteria optimization problem was reduced to a three-criteria optimization

problem where the D(5)-efficiency (based on the single point estimate of d from a pilot study) summarizes precision in model parameter estimation across the relevant range of the variance ratios.

The Pareto approach for design selection involves two stages. The first objective stage generates a set of Pareto optimal designs with the PAPESPD search algorithm. This set of designs is known as the Pareto optimal set and the corresponding criteria vectors (i.e., $\mathbf{F}(\xi) = (D(5),\ WP, N)'$ where ξ denotes a design in the Pareto optimal set) make up the Pareto front. In the second stage, this front is first reduced to a smaller, more manageable set and each design is evaluated over a mesh of weighting combinations of the three criteria of interest using the L_1-norm as a distance metric to the Utopia point. Though the examples discussed were based on specific design criteria, the methodology is easily adapted to other design criteria.

ABOUT THE AUTHORS

Dr. Lu Lu is a visiting assistant professor at the University of South Florida and was a postdoctoral research associate with the Statistical Sciences Group at Los Alamos National Laboratory. Her statistical research interests include reliability, design and analysis of experiments, response surface methodology, survey sampling, and analysis. She is a graduate of the Statistics Department at Iowa State University.

Dr. Timothy J. Robinson is the Director of the WWAMI Medical Education Program and a Professor in the Department of Statistics. He is a Fellow of the American Statistical Association and a Fellow of ASQ. His e-mail address is tjrobin@uwyo.edu.

Dr. Christine M. Anderson-Cook is a research scientist in the Statistical Sciences Group at Los Alamos National Laboratory. Her research interests include design of experiments, reliability, and response surface methodology. She is a Fellow of the American Statistical Association and the American Society for Quality.

REFERENCES

Anbari, F. T., Lucas, J. M. (1994). Super-efficient designs: How to run your experiment for higher efficiency and lower cost. *ASQC Technical Conference Transactions*, 852–863.

Arnouts, H., Goos, P. (2010). Update formulas for split-plot and block-designs. *Computational Statistics and Data Analysis*, 54: 3381–3391.

Bingham, D., Sitter, R. R. (1999). Minimum-aberration two-level fractional factorial split-plot designs. *Technometrics*, 41: 62–70.

Bingham, D., Sitter, R. R. (2001). Design issues for fractional factorial experiments. *Journal of Quality Technology*, 33: 2–15.

Bisgaard, S. (2000). The design and analysis of $2^{k-p} \times 2^{q-r}$ split-plot experiments. *Journal of Quality Technology*, 32: 39–56.

Box, G. E. P., Draper, N. R. (1975). A Basics for the selection of a response surface design. *Journal of the American Statistical Association*, 54: 622–654.

Cornell, J. (2002). *Experiments with Mixtures: Design, Models, and the Analysis of Mixture Data*, 3rd ed. New York: Wiley.

Derringer, G., Suich, R. (1980). Simultaneous optimization of several response variable. *Journal of Quality Technology*, 12: 214–219.

Goos, P. (2002). *The Optimal Design of Blocked and Split-Plot Experiments*. New York: Springer.

Goos, P., Vandebroek, M. (2001). Optimal split-plot designs. *Journal of Quality Technology*, 33(4): 436–450.

Goos, P., Vandebroek, M. (2003). D-optimal split-plot designs with given numbers and sizes of whole plots. *Technometrics*, 45: 235–245.

Goos, P., Vandebroek, M. (2004). Outperforming completely randomized designs. *Journal of Quality Technology*, 36(1): 12–26.

Gronwald, W., Hohm, T., Hoffmann, D. (2008). Evolutionary Pareto-optimization of stably folding peptides. *BMC-Bioinformatics*, 9: 109–121.

Huang, P., Chen, D., Voelkel, J. O. (1998). Minimum-aberration two-level split-plot designs. *Technometrics*, 40: 314–326.

Jones, B., Goos, P. (2007). A candidate-set-free algorithm for generating D-optimal split-plot designs. *Applied Statistics*, 56: 347–364.

Kasprzak, E. M., Lewis, K. E. (2001). Pareto analysis in multiobjective optimization using the collinearity theorem and scaling method. *Structural Multidisciplinary Optimization*, 22: 208–218.

Liang, L., Anderson-Cook, C. M., Robinson, T. J. (2006). Fraction of design space plots for split-plot designs. *Quality and Reliability Engineering International*, 22: 275–289.

Liang, L., Anderson-Cook, C. M., Robinson, T. J. (2007). Cost-penalized estimation and prediction evaluation for split-plot designs. *Quality and Reliability Engineering International*, 23: 577–596.

Liang, L., Anderson-Cook, C. M., Robinson, T. J., Myers, R. H. (2006). Three-dimensional variance dispersion graphs for split-plot designs. *Journal of Computational and Graphical Statistics*, 15: 757–778.

Lu, L., Anderson-Cook, C. M. (2012). Rethinking the optimal response surface design for a first-order model with two-factor interactions, when protecting against curvature. *Quality Engineering*, 24: 404–422.

Lu, L., Anderson-Cook, C. M. (2013). Adapting the hypervolume quality indicator to quantify trade-offs and search efficiency for multiple criteria decision-making using Pareto fronts. *Quality and Reliability Engineering International*, 29: 1117–1133.

Lu, L., Anderson-Cook, C. M. (2014). Balancing multiple criteria with a Pareto front for optimal split-plot designs. *Quality & Reliability Engineering International*, 30: 37–55.

Lu, L., Anderson-Cook, C. M., Robinson, T. J. (2011). Optimization of designed experiments based on multiple criteria utilizing Pareto frontier. *Technometrics*, 53: 353–365.

Lu, L., Anderson-Cook, C. M., Robinson, T. J. (2012). A Case study to demonstrate Pareto frontiers for selecting a best response surface design while simultaneously optimizing multiple criteria. *Applied Stochastic Models in Business and Industry*, 28: 206–221.

Myers, R. H., Montgomery, D. C., Anderson-Cook, C. M. (2009). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 3rd ed. New York: Wiley.

Park, Y. J. (2009). Multi-optimal designs for second-order response surface models. *Communications of the Korean Statistical Society*, 16: 195–208.

Parker, P. A., Anderson-Cook, C. M., Robinson, T. J., Liang, L. (2008). Robust split-plot designs. *Quality and Reliability Engineering International*, 24: 107–121.

Parker, P. A., Kowalski, S. M., Vining, G. G. (2006). Classes of split-plot response surface designs for equivalent estimation. *Quality & Reliability Engineering International*, 22: 291–305.

Robinson, T. J., Anderson-Cook, C. M. (2011). A closer look at D-optimality for screening designs. *Quality Engineering*, 23: 1–14.

Searle, S. R., Casella, G., McCulloch, C. E. (1992). *Variance Components*. New York: John Wiley & Sons.

Smucker, B. J., del Castillo, E., Rosenberger, J. L. (2011). Model-robust designs for split-plot experiments, *Computational Statistics and Data Analysis*, 56: 4111–4121.

Trinca, L. A., Gilmour, S. G. (2001). Multi-stratum response surface designs, *Technometrics*, 43: 25–33.

APPENDIX

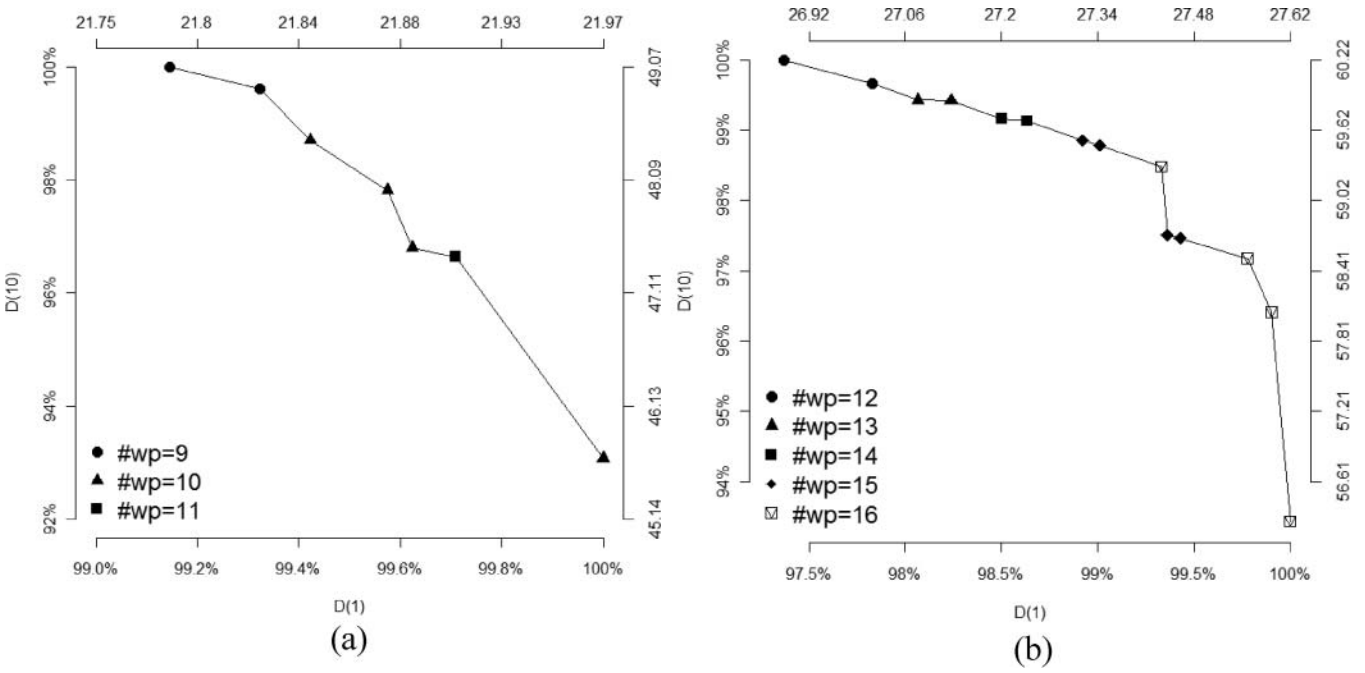

FIGURE A1 **Pareto front based on D(1) and D(10) for five continuous nonmixture factors (three whole plots and tw subplot factors) with a first-order model plus two-factor interactions for (a) $N = 20$ and (b) $N = 25$ with different *#WPs*.**

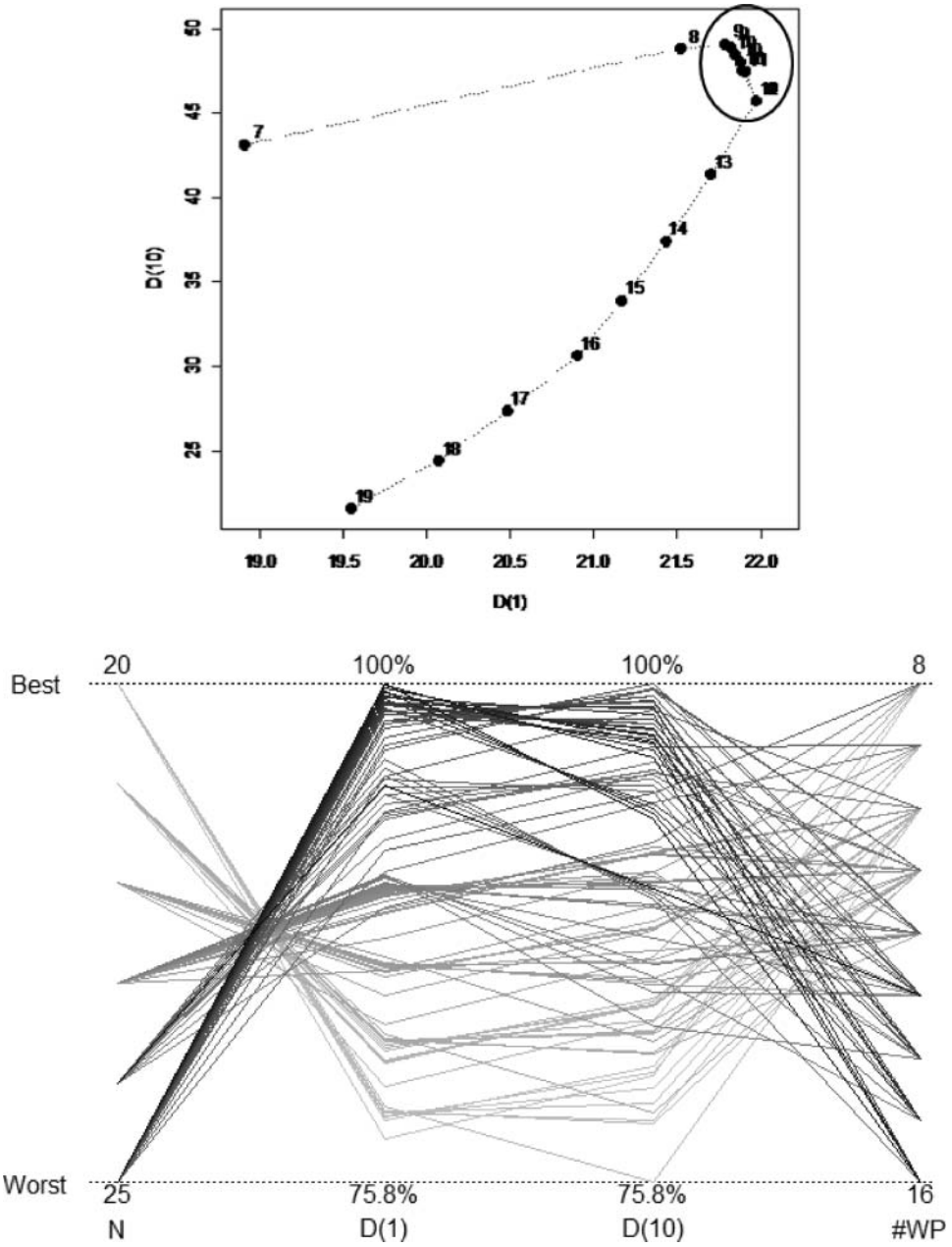

Figure A2 **(a) Pattern of best designs based on D(1) and D(10) for a nonmixture example for $N = 20$ and different *#WP*. The circled symbols in (a) correspond to the Pareto front based on the three criteria: D(1), D(10), and *#WP*. (b) Parallel plot of the overall four-criteria (D(1), D(10), *N*, *#WP*) Pareto front for the nonmixture example with $N = 20$–25.**