Received: 14 June 2017 | Revised: 5 February 2018 | Accepted: 21 February 2018

DOI: 10.1002/asmb.2325

DISCUSSION PAPER

WILEY

Graphics to facilitate informative discussion and team decision making

Christine M. Anderson-Cook[1] | Lu Lu[2]

[1]Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, NM, USA

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

Correspondence
Christine M. Anderson-Cook, Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, NM 87545 USA.
Email: c-and-cook@lanl.gov

Everyone knows the expression "A picture is worth a thousand words," and this effectively summarizes the ability of graphical summaries to convey information and persuade. However, in many cases, the goal for the right visualization is to encourage and guide discussion while helping focus a team to make carefully considered, defensible, and data-driven decisions. The aims of graphics differ if we are trying to communicate the merits of a single choice versus outlining several contending alternatives for further comparison and discussion. These choices each have their own strengths and weaknesses depending on how we value different criteria. They also serve different purposes at various stages of decision making. Often the role of statisticians is not to provide a single answer but to provide rich information and summaries in a manageable and compact form to enable productive discussion among team members. Through a series of diverse examples, we present principles and strategies for encouraging discussion and informed decision making and discuss how they can be integrated with versatile use of graphical tools for examining multiple objectives, framing trade-offs between alternatives, and examining the impact of subjective priorities and uncertainty on the final decision.

KEYWORDS

desirability function, DMRCS decision-making process, graphics for comparison, Pareto front, scalability to higher dimensions, trade-offs

1 | INTRODUCTION

A first experience with the power of graphics to initiate and facilitate a discussion came for one of the authors (Anderson-Cook), when presenting a summary of a new approach for predicting the reliability of a complex system to a group of reliability engineers and managers. The goal of the analysis was to determine if it was appropriate to certify the units in the stockpile (or population) as adequately reliable to remain in the field for additional 3 years (36 months). Hence, extrapolation to predict the reliability beyond the currently observed ages was intrinsic to the problem, and the summary plot provided to the decision team included the estimated reliability curve with associated uncertainty interval for each age, drawn from age 0 to 12 years (144 months) beyond the age of the oldest unit. Figure 1A shows a sample of the plot with the vertical dashed line indicating the current age of the oldest unit. Anderson-Cook had naively assumed that the part of the presentation that would lead to vigorous discussion was the detailed description of underlying statistics of the methods that used multiple sources of data, instead of just full system tests.[1,2] However, the simple reliability plot ignited a firestorm of discussion. Why? Because Anderson-Cook had extrapolated the estimated curve beyond the age of the oldest unit to include the ages about which a decision had to be made. Historically, the plots had just shown the reliability curve in the observed age range of the units, and each member of the team has individually and privately

 © 2018 John Wiley & Sons, Ltd.

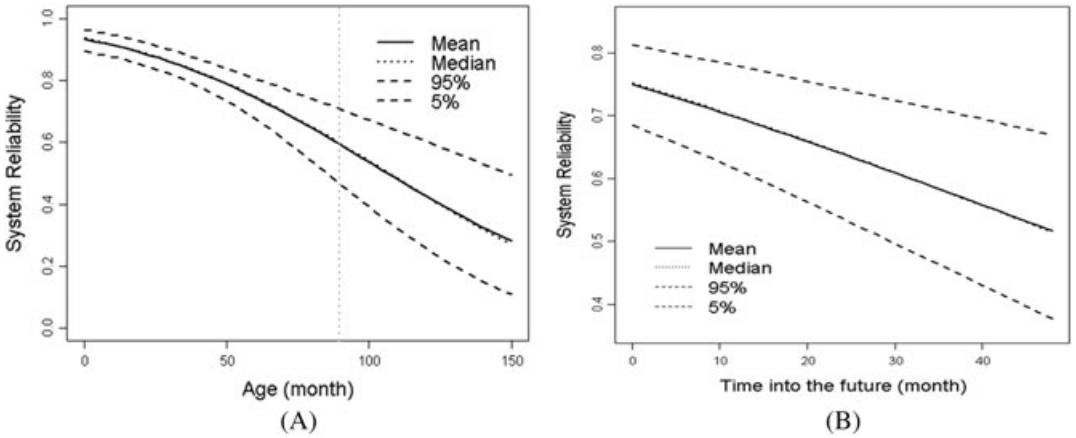

FIGURE 1 A, The individual reliability (IndRel) summary as a function of the age of the unit; B, The population reliability (PopRel) summary as a function of the time from present

extrapolated the curves to the decision point in their heads. With one plot, the discussion was changed, from individuals having their own private view of the decision space to one where the crux of the decision was now out in the open and available for direct discussion. Some team members thought the curve was too optimistic, whereas others thought it was too pessimistic. However, being presented with a concrete visualization of the analysis that directly connected to the decision at hand sharpened the focus of the discussions and positively changed the nature of the decision making. This seemingly obvious choice to show the quantity on which the decision will be made, namely, for this example to match the plotted age range to show the extrapolated reliability prediction, is the first of multiple strategies that are presented in this paper.

This example illustrates the use of graphics to effectively convey statistical analysis results and facilitate communication and discussion for making data-driven decisions. It is worth noting that graphics can play a broader role in the entire quantitative problem solving process, from the exploration stage to the formal analysis and to the final stage of communicating results and making decisions. The power of graphical summaries in exploratory data analysis and formal statistical analysis has been broadly discussed and emphasized through various applications. This paper focuses on the use of graphics in the decision-making process and how having the right graphics can support making transparent and informed evidence-based decisions.

In the remainder of this paper, we present 5 examples from our experiences with facilitating discussion and decision making that illustrate some strategies for making effective use of graphics with this goal. Rather than focusing on graphics that are persuasive toward a particular conclusion or exploratory to discover new aspects of a set of data, in this paper, we highlight approaches for comparing alternatives to encourage discussion and incorporating subjective prioritizations that different decision makers bring to the process. The goal is to use carefully designed graphical tools to support informed discussion on considering alternatives and helping reach consensus on a more justifiable decision. The remainder of the introduction connects the roles of statisticians as facilitators to the broader discussion of these graphical strategies and describes the define-measure-reduce-combine-select (DMRCS) process[3] for decision making. Section 2 revisits the reliability example described to highlight other aspects of the decision. Section 3 illustrates a design selection example where graphical summaries help examine multiple aspects of the design performance for comparing several competing choices to justify a final decision. Section 4 examines approaches to choosing an optimal designed experiment when there are competing objectives. Section 5 highlights the roles of variability and uncertainty in the decision-making process and how graphical summaries can consider these aspects when optimizing multiple responses. Section 6 considers how to allocate resources as part of stockpile prioritization where multiple diverse aspects need to be balanced. Finally, the conclusions in Section 7 recap the general strategies outlined and provide some suggestions for implementing these ideas for other applications.

In many business, industries, and government settings, statisticians are finding opportunities to participate in decision-making teams. Anderson-Cook[4] discusses several keys to success when contributing statistical thinking and uncertainty quantification to the process. These suggestions include translating the problem into a discussion of trade-offs based on quantitative summaries of the objectives, streamlining the choices through elimination of noncompetitors, and providing visualization tools to facilitate discussion. It is this third idea that we wish to expand upon with some concrete suggestions for how to think about which graphics are beneficial and how to present them.

The ideas in this paper complement the DMRCS structured decision-making framework suggested in the work of Anderson-Cook and Lu.[3] This framework focuses on translating an often vaguely defined set of aspects of a decision and transforming them into specific measurable metrics, which can then be compared and contrasted across a range of different prioritizations from different members of the decision-making team. The process culminates with a final decision that

is understood in the context of the alternatives. Similar in spirit to the 6 sigma define-measure-analyze-improve-control (DMAIC) (see the work of Hoerl and Snee[5,p. 128-137]), the decision-making counterpart DMRCS provides a structure for selecting the best-suited choice based on several likely competing objectives. The Define step focuses on clarifying the parameters of the decision to be made, characteristics to consider, and the space of possible solutions. As with DMAIC, this step is critical for framing the decision to ensure a relevant solution is found. It is important to think broadly about diverse aspects relevant to the decision and represent all key facets with an appropriate metric.

Another step that overlaps with the DMAIC process is the Measure step, which emphasizes having high-quality data on which to base the decision. The quantitative criteria should be precisely and accurately measured to allow fair and consistent comparisons between potential solutions. The goal of the Reduce step is to simplify and streamline. First, optimizing based on too many criteria often leads to mediocre results individually, as the trade-offs between criteria typically increase with the addition of more criteria. The second type of reduction to consider is eliminating noncontending solutions from further consideration with the construction of a Pareto front (PF).[6] Understanding the highlighted choices is an opportunity for graphical summaries.

The Combine step looks at information from different criteria, often measured on different scales simultaneously, and considers graphical methods for representing different facets of the decision. Here, the decision makers' priorities and how much they value good performance on different criteria are essential to identifying the best and most relevant solutions. Finally, the Select step identifies the available solution best suited to the decision makers' objectives and allows comparisons between close contenders. At the conclusion of the process, the decision-making team should have a preferred choice and be able to articulate why it is best given their priorities.

As a reviewer pointed out, the DMRCS decision-making process is not completely independent from the DMAIC problem-solving model but could be a useful complement when used for guiding an improved decision. In the DMAIC context, decision making is involved when solutions are developed for dealing with the root causes identified from the "Analyze" step, which are then tested for validity in the "Improve" step. The DMRCS model can add structure and rigor to the often complicated and unstructured decision-making process and provides a better decision than what is often used in the DMAIC process.

Table 1 contains a list of 15 recommendations about how to conceptualize, construct, and use graphical methods to assist with the structured decision-making process. We have organized the list into categories that include general suggestions and for each of the steps in the DMRCS process. The individual suggestions are highlighted in one or more of the examples throughout the rest of this paper. We now consider the examples and use them as vehicles for outlining strategies to use graphics to help facilitate discussion and decision making.

TABLE 1 The general and define-measure-reduce-combine-select principles for improved discussion and decision making

	General
G1	Use a process to find common ground
G2	Clarify the difference between "right and wrong" and "choosing differently based on priorities"
	Define
D1	Define and use an appropriate summary that directly connects to the decision
D2	Build the decision space to include diverse alternatives
D3	Push the boundaries on assumptions
D4	Incorporate cost into the comparisons of alternatives
	Measure
M1	For realistic decision making, it is important to provide uncertainty quantification to inform the decision makers about the uncertainty and potential risks associated with a specific decision
M2	Devise analysis summaries that reduce the uncertainty can improve the decision-making process
	Reduce
R1	It is helpful to take strategic steps to reduce the number of choices on which to do a detailed comparison to a manageable number
	Combine
C1	Allow common visualization and discussion about results, instead of keeping subjective elements of the decision unshared
C2	When comparing the smaller set of contenders, choose plots that highlight the impact of subjective choices to facilitate discussion
C3	Think globally and locally
C4	Consider dynamic graphics when dimensionality of problem suggests it
	Select
S1	Create or use a graphical summary that appropriately captures the needed level of detail
S2	Include graphics to formalize conclusions

2 | RELIABILITY ASSESSMENT FOR POPULATIONS OF UNITS

In the introduction, an initial description of the reliability assessment decision was outlined. We now provide some additional details and how graphical tools helped to facilitate an improved final outcome. The reliability of the individual units can be assessed most directly from full system tests, which are both destructive and expensive. As a result, using exclusively these data often leads to a wide uncertainty interval about the estimated or projected reliability (if a prediction into the future is needed). To help reduce this uncertainty, several strategies were used, including using results from nondestructive component and subsystem level tests.[1,2] In addition, the analysis incorporated subject matter expertise leveraged from design specifications and available data from sister systems sharing common characteristics and components. This information was incorporated into the analysis through the use of informative priors with a Bayesian analysis.[7]

The plot of the estimated reliability of individual units projected to the age range of interest in Figure 1A was key to initiating discussion and helping the decision-making team focus on data that was directly relevant to the decision. While an improvement over previous figures that did not include all of the age range of interest, this was still not the right summary for the decision. After further discussion, it became clear that there was still a gap between the information in Figure 1A and what was needed for the final decision. The stockpile was comprised of units of various ages, since they were purchased and added to the population over multiple years. The standard that dictated the certification process was based on a lower uncertainty bound for the overall reliability of the population. Hence, while helpful to understand the pattern of reliability changing as a function of age, the actual question of interest required a different summary. Lu and Anderson-Cook[8] define the Population Reliability (PopRel) summary that calculates the probability of a randomly selected unit from the population working successfully at a current or future time point. In comparison with the individual reliability (IndRel) summary as a function of age of the units shown in Figure 1A, the PopRel summary as a function of the time from present is shown in Figure 1B, and this is a direct match to the requirements for the decision regarding the stockpile of the units. Hence, to make the best possible decision, a number of factors contributed to the decision-making process.

D1. Define and use an appropriate summary that directly connects to the decision (here, using PopRel, instead of looking at the IndRel). Seemingly obvious, it is surprisingly common to try to use a convenient and yet indirect summary for decision making, when with a bit of additional effort, a direct summary can be constructed.

M1. For realistic decision making, it is important to provide uncertainty quantification to inform the decision makers about the uncertainty and potential risks associated with a specific decision. The uncertainty bounds based on the Bayesian credible interval are effective and informative to show the range of possible performance consistent with observed data, instead of just the average performance (point estimate based on the mean). This allows an appropriate and realistic decision based on what is actually known about the reliability.

M2. Devise analysis summaries that reduce the uncertainty can improve the decision-making process. It is helpful to include supplementary but relevant data to better inform the decision (here, the combined analysis using component and subsystem level data improved the analysis and made the credible intervals narrower). Use whatever is available and relevant to get as much information included in the decision-making process. In addition, when possible incorporate subject matter expertise (here, a Bayesian analysis was used to allow for external sources of information to be included through the use of informative priors). Digging more deeply to include additional information allows for more complete assessment and deepens the buy-in from the decision-making team.

C1. Allow common visualization and discussion about results, instead of keeping subjective elements of the decision unshared (here, extrapolating reliability results to include the time point for the decision). When the subjective component and its impact are not available for examination and discussion, the decision-making team is likely to struggle to understand each other's perceptions and positions.

3 | SELECTING AN IDEAL DESIGNED EXPERIMENT BASED ON MULTIPLE CHARACTERISTICS

In this section, we consider a commonly occurring scenario of how to select the best designed experiment for a particular set of objectives. Anderson-Cook and Lu[9,10] considered this multifaceted decision for a scenario when they were asked to suggest a 14-run design in a design space where relatively little was known about the underlying relationship between input factors and the response. Despite the original request for the statisticians to provide the design, we encourage

providing several carefully selected alternatives that help to facilitate a discussion about what the goals are for the experiment and which of several alternatives makes the most sense to best match these goals.

For this example,[9] the original request from the engineer was to suggest a 14-run design for a screening experiment involving 7 factors with a primary goal of estimating the main effects model with some concerns about potential active 2-factor interactions. However, 4 potential designs were examined in the end including a 14-run D-optimal design (14r), the same design with a center run added (15rCR), a 15 run definitive screening design[11] (15DSD), and a 16-run D-optimal design (16r). An initial reaction to these choices might be that the statisticians were not very good at following instructions, as all but one of the designs do not match the original request. While this is true, an important aspect of good decision making is to probe the edges of the assumptions. In this case, the cost of the experiment (as defined by the number of runs available to stay within budget) can be examined. Understanding what a slightly larger (or smaller) design can offer can spawn discussion about whether the requested design size is adequate for the intended purpose and how that size was originally determined. If the budget is truly fixed with no latitude for change, then it can still be beneficial to understand its performance, both good and bad, in the context of alternatives. Since designed experiments are often run as part of a larger sequence of exploration, finding the appropriate size should be examined in this context, ie, can we learn enough from this experiment to be well-positioned for the next planned step? Or would a smaller design allow enough information to be learned while saving more resources to be available at later stages in the sequence?

Considering alternatives frames the choices in a way that encourages questioning. For example, if you are shopping for dish soap, the process looks very different if you are faced with near-empty shelves compared with having multiple choices. If there is only a single choice, mentally, you are trying to justify that this choice could work. With several possibilities, different alternatives are weighted and comparisons between the choices are natural (Is the extra cost worth it for name brand? How big a bottle is needed? What scent do I really want?). For the dish soap, it is likely that there are not too many bad choices, and the stakes are low. However, for a designed experiment, there likely are many quite poor choices that can impact the quality of results and path forward. Therefore, considering different options is important, but what aspects of the design should be compared to make a good choice?

A clear sense of what the designed experiment is trying to accomplish is important. For this experiment, the primary goal was to gain understanding about which of the 7 inputs factors and their interactions are most influential on the response. Hence, good estimation of the model parameters was paramount. Both the ability to formally test if different sized effects are statistically significant, and avoiding ambiguity from confounding with other parameters are important. Figure 2 shows several graphical summaries that compare multiple design choices based on relevant criteria. Figure 2A shows the power[12] of each design for all of the main effects as a function of how large the true effect is relative to the natural variability (σ) of the experiment. For example, if an effect has size 0.5σ, then the 15DSD has the worst power, with about a 30% chance of finding it statistically significant, whereas the 16r design has about a 45% chance. Figure 2C shows the map of correlations between different model parameters (main effects and 2-factor interactions). Ideally, the correlation would be zero between all terms (denoted in blue), but for small designs that are looking to estimate a larger model, this can often not be achieved. In these cases, it is of interest to examine how large the correlations are and between which terms in the model (confounding effects). As can be seen in the figure, the designs have quite different correlation patterns, ranges, and model terms affected. For example, the 15 DSD has no main effects correlated with any 2-factor interactions, whereas 16r D-optimal design has completely unconfounded main effects.

A secondary objective of the experiment is to be able to predict the response well throughout the design space. The fraction of design space plot[13] shown in Figure 2B allows direct comparisons between alternative designs, where consistent low prediction variance (flat curve with small values) across all locations in the design space is ideal. Anderson-Cook and Lu[9,10] discussed other aspects of the designs that are also important to consider, such as the capability of assessing curvature and other optimality criteria that emphasize on different aspects of the design characteristics.

When comparing potential designed experiments, examining multiple aspects of the designs is highly desirable. Considering both the primary and secondary goals of the experiment and probing the outcome depending on different assumptions is beneficial. Note how both Figures 2A and 2B are compact summaries that allow for multiple choices to be directly compared. When possible, this helps with easy to understand summaries of an important characteristic of interest. Alternately, Figure 2C shows detailed information that could be lost if reduced to low-dimensional summaries. In this case, it is important to be able to look at this rich set of interest that has been organized to ease interpretability. In this summary from JMP software,[14] the main effects are all listed first, which allows easy examination of main effect–by–main effect, main-by-interaction, and interaction-by-interaction correlations (shown in different blocks of regions in the correlation map). While a great deal of information is available in the plot, it can be examined at a high level to see gross characteristics of each design.

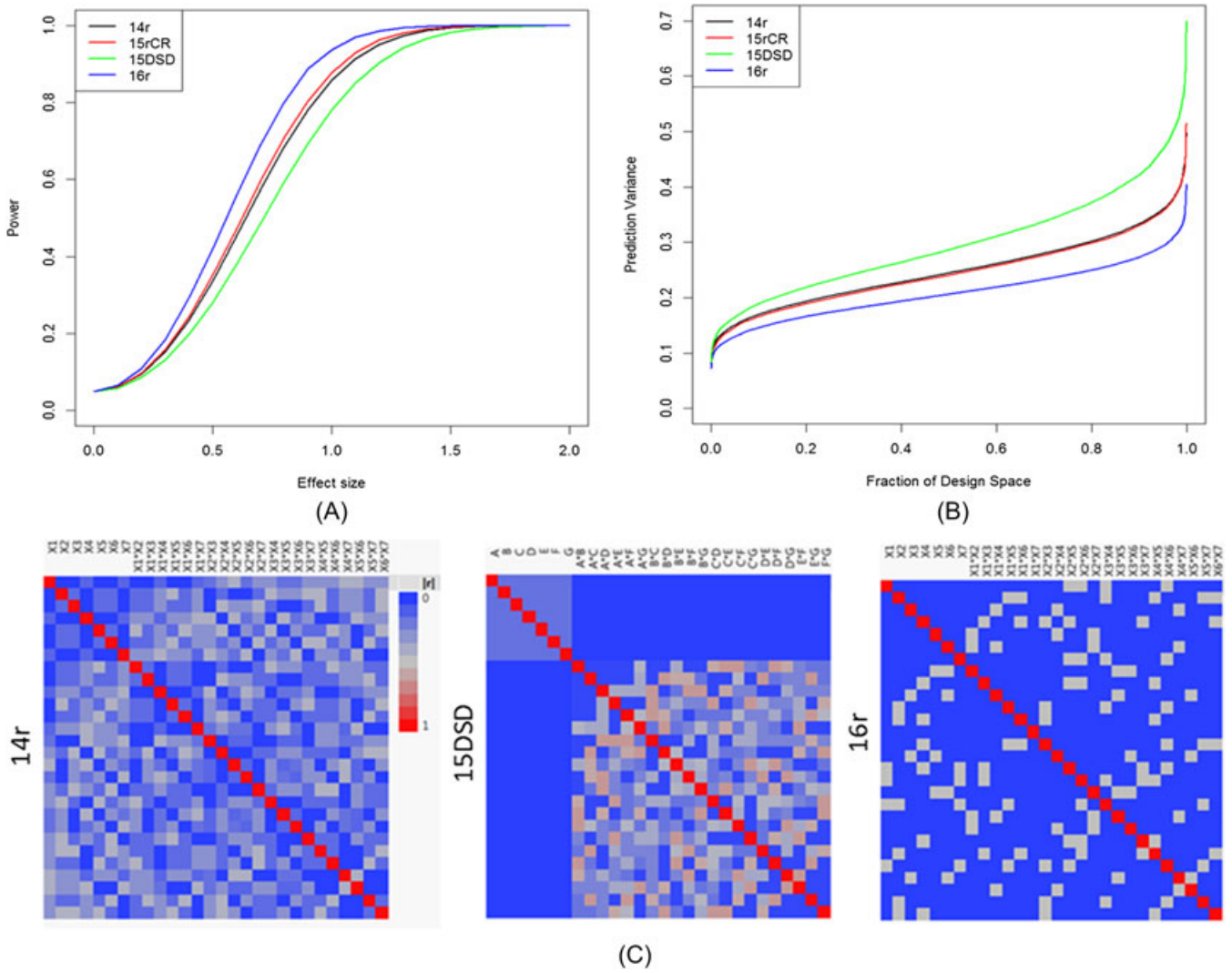

FIGURE 2 Graphical tools for design diagnostics. A, Power summary: estimated probability of finding any main effect of a given size statistically significant at the 5% significance level (ie, *p*-values less than 0.05) for the 4 designs (14r, 15rCR, 15DSD, and 16r) discussed in Section 3; B, Fraction of design space plot to show the quantiles of the prediction variance across the design region for factor level values in [−1,+1] for the 4 designs; C, The correlation color maps for 14r, 15DSD, and 16r designs [Colour figure can be viewed at wileyonlinelibrary.com]

We now summarize some of the lessons from this scenario.

G2. Clarify the difference between "right and wrong" and "choosing differently based on priorities" (here, different team members are likely place different emphasis on some of the criteria or worry to different degrees about what might go wrong. This should be a basis for discussion, not to be critical of others' opinions).

D2. Build the decision space to include diverse alternatives (here, considering larger designs than formally requested facilitated deeper probing of what is needed and whether the original choice on the design size is appropriate for design goals).

D3. Push the boundaries on assumptions (here, the inclusion of a rich model including interactions provides deeper understanding about what the outcome of the experiment might involve). An important part of the process is articulating assumptions, and, since these are often based on imperfect understanding, it is helpful to consider what the outcome will be depending on how different assumptions turn out.

D4. Incorporate cost into the comparisons of alternatives (here, expanding the mandate of the experiment to look at alternatives of different size helps understand the space of options). Since cost (in time, effort or dollars) enters into almost all decisions, formally including this in the decision-making process is important to fully explore the space. Plotting the performance of different priced experiments against each other encourages discussion of what solution is best for what is needed.

S1. Create or use a graphical summary that appropriately captures the needed level of detail. When possible, direct comparisons between alternative solutions are desirable, but for some rich criteria with richer dimensions of information, the most desirable summaries allow for both high-level patterns and individual details to be extracted.

4 | PARETO FRONTS FOR DESIGN OPTIMIZATION

We now consider another related design of experiments problem where, in this case, the goal is to create a design for specialized scenario, where standard designs available in most statistical software does not necessarily match the aims of the experiment. A variety of different scenarios with nonstandard goals[6,15,16] provide examples of how different design metrics can be selected to emphasize different priorities. For the split-plot designs, the performance of the design, as measured by D-optimality for good model parameter estimation, is dependent on the ratio of the whole plot (WP) to subplot (SP) variances ($r = \sigma_{WP}/\sigma_{SP}$).[17] Since the size of either of these variances is typically not known before the experiment is run, it is beneficial to select a design that will perform well across the range of the anticipated values. For the example considered in the work of Lu and Anderson-Cook,[18] the engineers familiar with the process had little idea about the value of the variance ratio, and so decided that the likely values of r could be in a very wide range [0.1,10]. The assumed model for the 2 WP and 1 SP factors was a first-order model with 2-factor interactions, and the design size was constrained to a maximum size of $N = 16$ runs. Each run was quite expensive to analyze, so there was interest in potentially using fewer than the maximum allowable design size ($N \in [12, 16]$). Resetting the hard-to-change factors for each WP is time intensive, so there was interest in limiting the number of WPs. Hence, the goal of the experiment was to construct an ideal design that optimized over 4 criteria, ie, (i) maximize D(0.1): D-optimality when $r = 0.1$, (ii) maximize D(10): D-optimality when $r = 10$, (iii) minimize N: the overall design size, and (iv) minimize #WP: the number of WPs.

To find a set of contending designs to compare, the Pareto aggregating point exchange for split-plot designs search algorithm[18] was used to explore the vast set of potential designs. Central to this exploration of candidate designs is the use of a PF, which considers only those solutions, here designs, that are not *dominated* by any other solutions. One solution S_1 is defined to dominate another solution S_2 if it has criterion values that are at least as good as S_2 for all criteria and is strictly better for at least one criterion. It can be shown that, for a chosen set of criteria, the set of solutions on the PF is the set of rational choices and will contain the best solution for any user priority often represented by the desirability function[19] weighting of the different criteria. Figure 3 shows the set of PFs for each of the possible design sizes, $N \in [12, 16]$ with #WP $\in [5, N]$ where the different #WP are noted with different symbols. Note this single plot shows all of

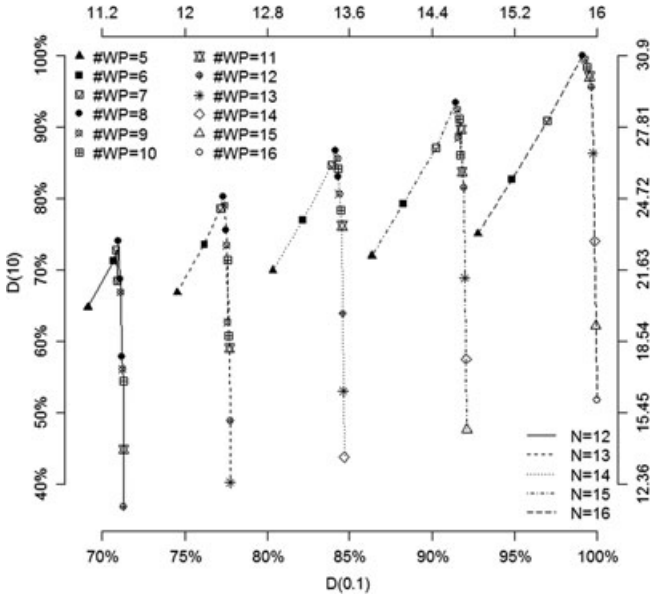

FIGURE 3 The paired values of (D(0.1) and D(10)) for all designs on the Pareto fronts of fixed combinations of N and #WP for an example with $N = 12$–16, and a first-order model with 2-factor interactions

the nondominated solutions based on the 4 criteria (D(0.1), D(10), *N*, and #WP). The D-optimality values are presented as relative efficiencies compared with the best possible design for each of D(0.1) and D(10).

From Figure 3, we see that the PF approach was able to distill the set of desirable options for design down to a manageable number of choices, with clear patterns in the results to help us understand the trade-offs between criteria. If we look at the PF for a single *N*, we see that for small #WP, increasing the number of WPs improves both D(0.1) and D(10). After reaching the maximum possible value of D(10) at #WP = 8 for a given *N*, further increasing the #WP leads to slight improvements to D(0.1) but substantial decreases in performance for D(10). By comparing across different PFs for the various *N*, we see the actual amount of benefits of increasing sample size for both D(0.1) and D(10).

When the engineers were presented with this summary, they were able to directly assess the cost benefit analysis of increasing or decreasing the sample size relative to the improvement in precision of the model parameters. Based on this, it was decided that the cost of using a design size of 16 was justified by the improvement in D-optimality across the range of anticipated *r* values. In addition, the engineers could universally agree that the improvement in D(0.1) beyond #WP = 8 was so minimal that it did not justify the sharp decrease in D(10) plus the additional time to run the experiment with an increased #WP. In this case, using the right summary in the Reduce step of the DMRCS process[3] with a manageable number of designs to consider and clear patterns made consensus among the decision-making team straightforward.

However, there was still a final decision to be made between the $N = 16$ designs. In this case, there was no universal agreement about how to value the differences in performance relative to the savings in time. Some additional graphical tools were needed to help understand the trade-offs. Figure 4 shows the mixture and trade-off plots[6] comparing these choices. The mixture plot (Figure 4A) shows what the best design is for all possible desirability function weightings of the 3 criteria (D(0.1), D(10), and #WP) using the L_1-norm on the log scale (which is equivalent to using the multiplicative desirability function), where the scaling from [0,1] is based on the worst and best available values from among the designs on the PF. Clearly when #WP is weighted very heavily (at the top of the simplex), the best design has #WP = 5. In the bottom left corner, there are some tiny slivers of weights for which the improvement in the D(0.1) values make the designs with #WP < 8 best. However, for a large fraction of the overall weights where the D-optimality criteria are weighted no less than 30%, the #WP = 8 design is best. In Figure 4B, the trade-off plot shows the range of performance values for the designs scaled by the desirability function scores in [0,1]. Here, it is possible to see the peak in D(10) for #WP = 8, and the monotonic decrease in D(10) as #WP decreases. From Figure 4, the decision-making team was able to identify their 2 top choices as $N = 16$ with #WP = 5 or 8. The use of effective graphical tools in the Combine step of DMRCS[3] had further reduced the set of promising solutions from 50 on the PF down to only 2 top choices from which to select.

Since only 1 design can ultimately be run, a final choice needed to be made. Figure 5 shows a direct comparison of performance of the these designs with some comparative summaries of how these designs compare relative to the best available designs at any weight combination of the criteria. The synthesized efficiency plot[15] in Figure 5A highlights the performance of a given design where white indicates that a design is between 95% and 100% efficient relative to the best available design considered. Each darker shade indicates a 5% drop in relative performance compared with the ideal.

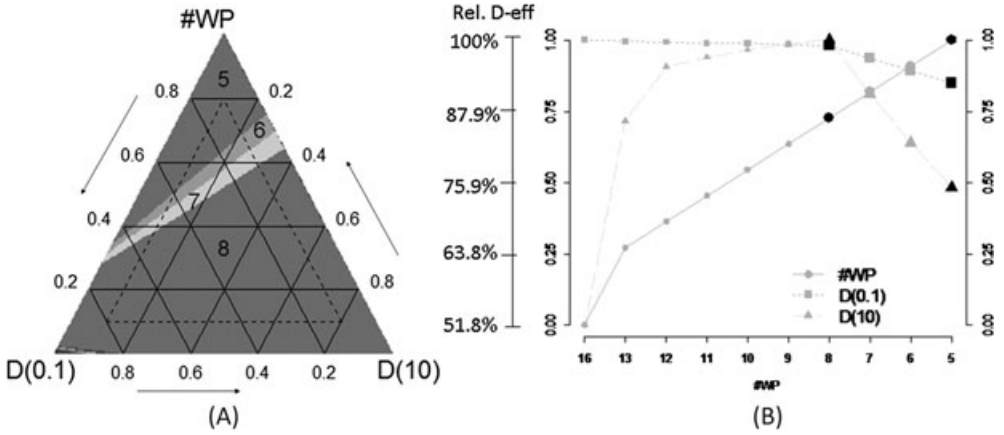

FIGURE 4 A, The mixture plot for designs selected from the Pareto front with $N = 16$, as shown in Figure 2, based on using the L_1-norm on the log scale for the 3 criteria, ie, D(0.1), D(10), and #WP. Designs with at least 10% weight for all 3 criteria and at least 1% of the total simplex area are highlighted with #WP shown in corresponding regions; B, The trade-off plot with black symbols for highlighted designs for designs shown in Figure 4A

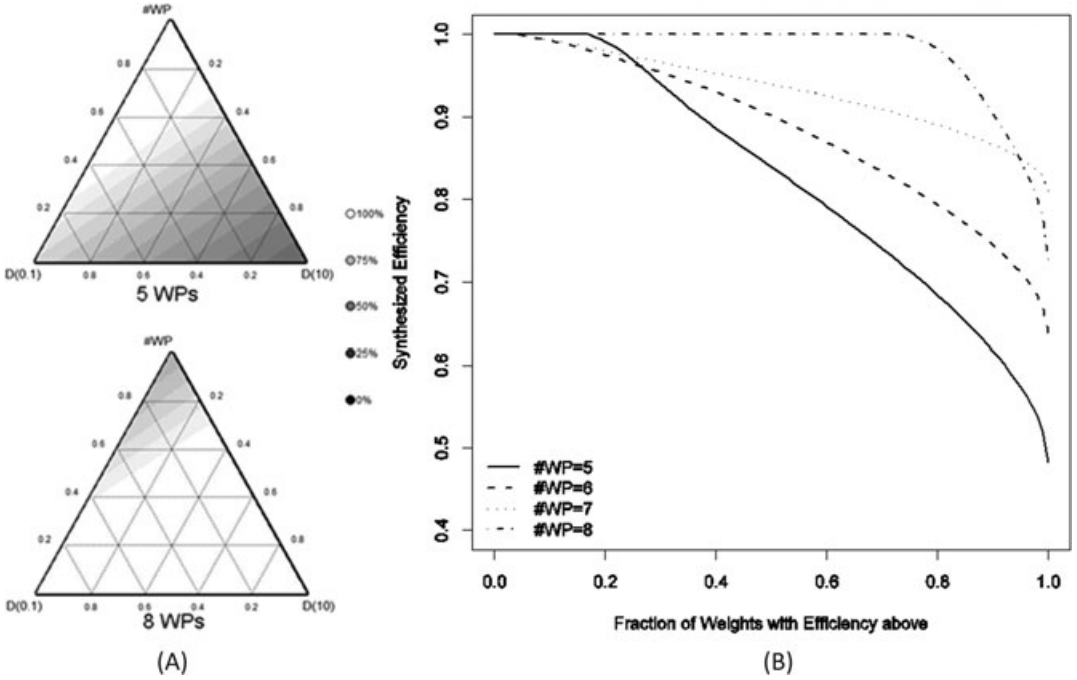

FIGURE 5 The synthesized efficiency plot of some selected designs (with 5 to 8 WPs) for different weightings of the 3 criteria, ie, D(0.1), D(10), and *#WP* based on the L_1-norm on the log scale for the 16-run split-plot design. A, Synthesized efficiency plot; B, Fraction of weight space plot

When we compare the #WP = 5 and 8 designs, we see that the decrement in performance for the #WP = 8 design is never below 75% efficient for any weight combination. The #WP = 5 design has some considerably darker shades for relative efficiency. The fraction of weight space (FWS) plot[20] in Figure 5B is a global summary of all of the scores across all of the weights. For a given point on the curve, it is possible to see what fractions of the weights (shown on the *x*-axis) have efficiency at least as large as the value on the *y*-axis. The ideal line on the FWS has high values across all possible weights. For this example, the #WP = 8 design stays near 100% large for almost 80% of the weights. After some deliberation in the final Select step of the DMRCS process,[3] the decision-making team was able to reach an agreement that the best design for their experiment was the $N = 16$ with #WP = 8 design. If the decision is not as straightforward as for this scenario, Lu et al[21] suggests an approach for summarizing just a portion of the total weight space that can be agreed upon among all team members. For example, it might be that the team can agree that no criterion should be weighted less than 20%. In this case, another version of the mixture, synthesized efficiency, and FWS plots can be constructed to just reflect this subregion. Looking at a more focused region can help build consensus on the team by identifying common areas of prioritization and also eliminate some unappealing potential solutions.

We now highlight some of the lessons from this example.

D1. It is important to define and use an appropriate summary that directly connects to the decision. In this case, considerable effort and care was taken by the engineers to select the right experiment specific metrics that characterized their priorities. The process of identifying these metrics before any individual designs were considered helpful to build team cohesiveness and focus the discussion. This initial identification of the metrics allowed for the Pareto aggregating point exchange for split-plot designs search algorithm to be implemented for seeking tailored solutions. Efficient graphical summaries were used to include all 4 of the metrics on a single plot to see how they were related to each other.

D2. Build the decision space to include diverse alternatives. If a single rough guess of the variance ratio was inserted to simplify the decision at the beginning, then it is very likely to choose a suboptimal solution if the initial guess is somewhat off from the unknown true value. By considering different possible WP to SP variance ratios, more realistic assessment of what is possible was included in the search for a best design. Having an appropriate sense of the plausible outcomes and how the designs will perform across that space is important.

D4. Incorporate cost into the comparisons of alternatives. Although the budget was originally set at 16 runs, it was helpful to explore whether this was the right size of experiment, or whether it would be possible to save some of the

resources for other priorities. In the end, the engineers chose to use the full budget but with greater understanding of why it was the right thing. The time to run the experiment was included in the cost discussion with different #WPs considered. A plot that compared the impact on performance relative to the different costs highlighted this trade-off.

R1. It is helpful to take strategic steps to reduce the number of choices on which to do a detailed comparison to a manageable number. The use of a PF objectively eliminated inferior solutions, allowed for the results of the search to be efficiently presented, and kept the team from being overwhelmed with too much information. From the graphical summary of the PF, systematic pattern could easily be extracted, and the priorities of the decision-making team were clarified.

G2. Clarify the difference between "right and wrong" and "choosing differently based on priorities (revisited)". Removing a noncontender from consideration based on agreed criteria is an objective choice, whereas preferring 1 PF solution to another is typically a matter of different priorities.

C2. When comparing the smaller set of contenders, choose plots that highlight the impact of subjective choices to facilitate discussion. The mixture, trade-off, and synthesized efficiency plots all allow direct comparisons of alternative solutions in different ways. With this suite of graphics, discussion about how different team members prioritize the criteria is straightforward, and the impact of different priories is clearly shown.

5 | PARETO FRONTS FOR SIMULTANEOUSLY OPTIMIZING MULTIPLE RESPONSES

In Section 4, the power of the PF to eliminate noncontenders to improve the focus for team discussion and decision making was illustrated in the context of designed experiments. In that scenario, the metrics for each design could be calculated with no associated uncertainty. In this next paradigm, we consider another situation where the use of the PF is very effective, but now, uncertainty from the natural variability of the response and the model fitting process associated with each result should be included when making a decision. We consider the example described in the works of Chapman et al[22] and Myers et al,[23] where the goal is to simultaneously optimize 3 responses from a chemical process by selecting a preferred location in a 2-dimensional (time and temperature) input space. Figure 6A shows overlaid contour plots of the 3 responses, where the goal is to maximize yield (solid lines), minimize molecular weight (dashed lines), and hit a target value for viscosity (shading indicates distance from the target) (See the works of Chapman et al[22] and Myers et al[23p. 333]).

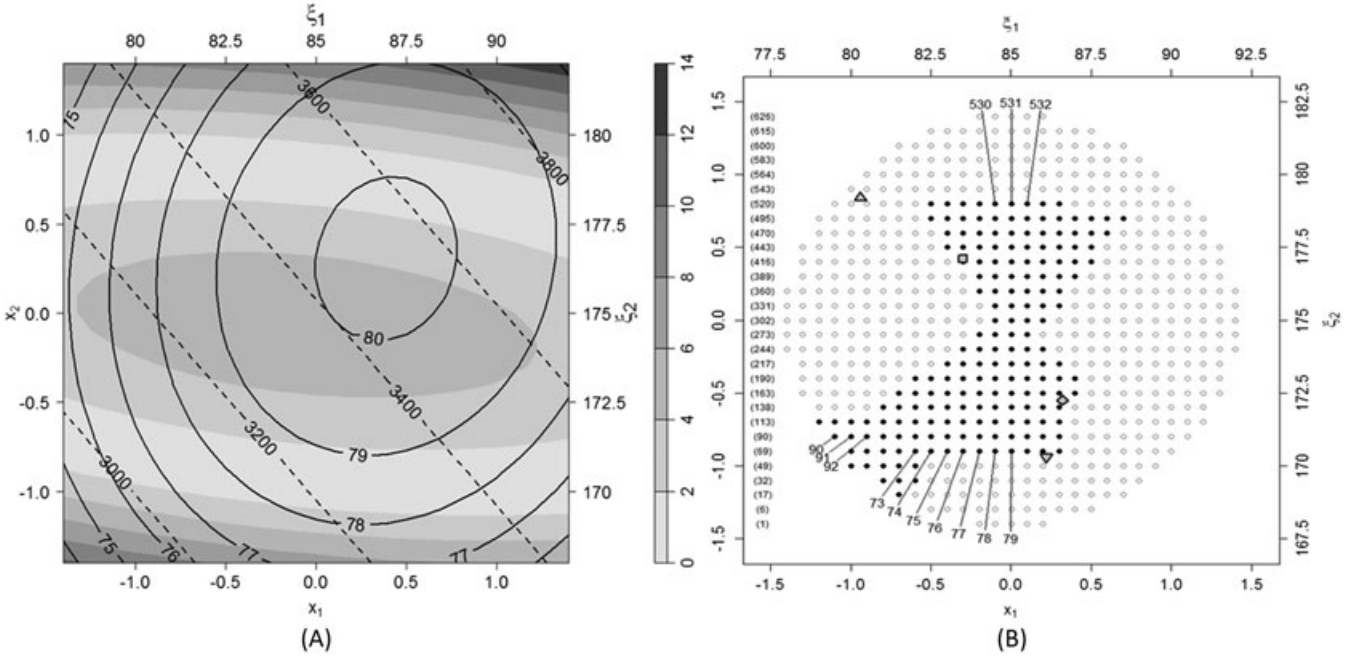

FIGURE 6 A, The overlaid contour plot for the estimated response surfaces for the example in the work of Myers et al[23]; B, The plot of input locations on the Pareto front (shown with the solid circles) based on simultaneously maximizing yield, minimizing molecular weight, and targeting viscosity at 65

A quick inspection of the 3 response surfaces shows that the individual optima lie in different regions of the design space, so compromise and balancing of trade-offs is required to find the best input factor combination for the process.

To find a set of locations to examine further, the 3 response surfaces were initially calculated with the mean model (using the point estimates of model parameters) at a grid of locations in the input space. The fineness of the grid was determined by the experimenters based on the precision with which they could set the time and temperature in the experiment. The filled circles in Figure 6B show the locations in the input space for which response values lie on the PF. Already, before any

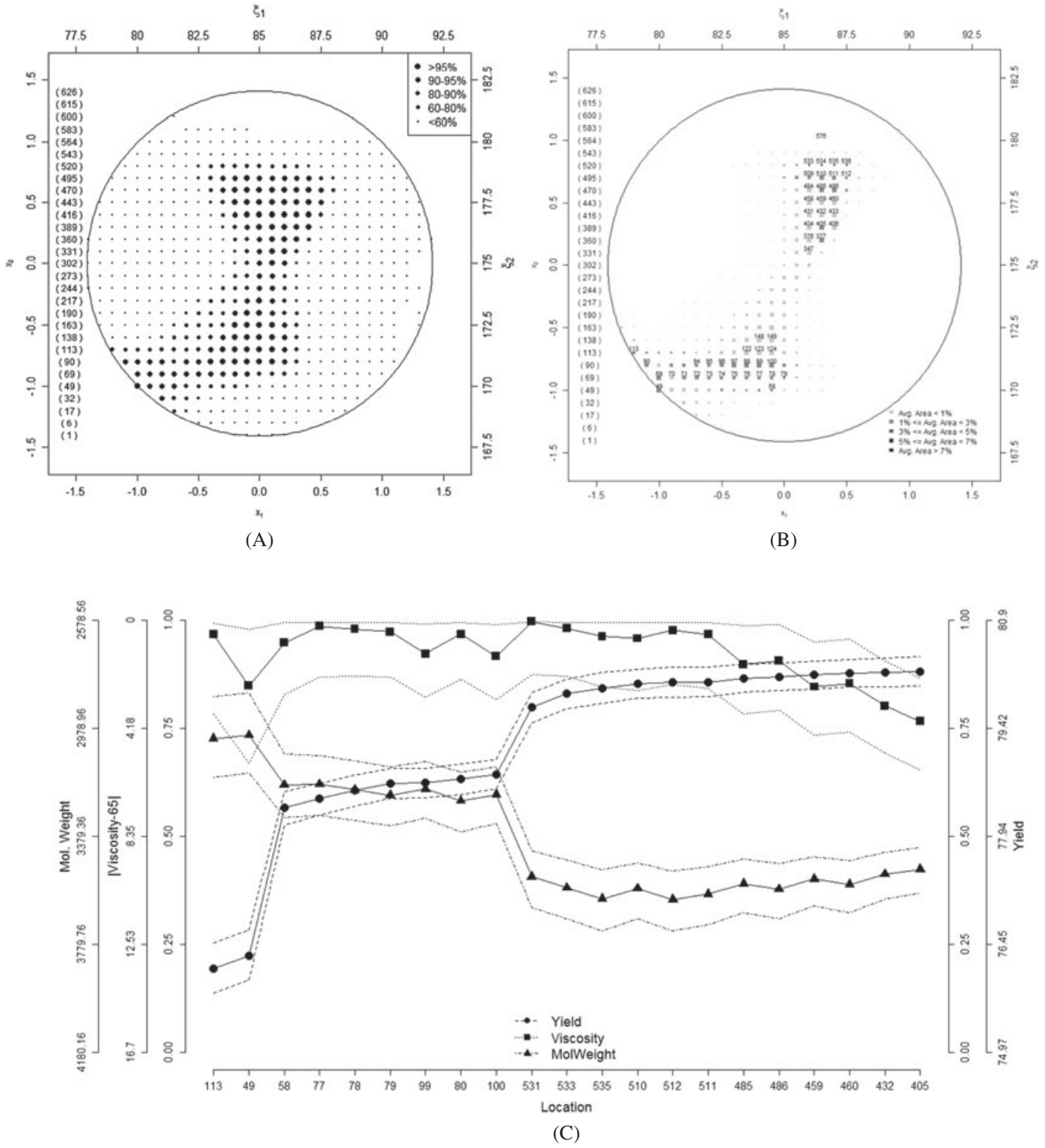

FIGURE 7 A, Plot of frequency of appearance on the Pareto front (PF) for input locations in the design space; B, The proportion of times a design location is best for at least one weight (size of symbol) and the average multiplicative mixture areas when a location is best (grayscale); C, Trade-off plot for the scaled responses with their uncertainty for promising solutions identified from the PF

prioritization of the 3 responses has been introduced, some regions of the input space (including only open circles) seem less desirable for the optimization. However, before eliminating these regions from further consideration, it is helpful to acknowledge that the uncertainty in estimation of the response surfaces has not yet been considered. In the work of Chapman et al,[22] 1 approach for incorporating this uncertainty is to look at a worst-case assessment of the responses from the 95% prediction intervals (P.I.). Hence, when the goal is to maximize yield, the lower bound of the 95% P.I. is used to represent the worst-case scenario. When minimizing molecular weight, the upper bound of the P.I. is used, and when we want to hit a target value for viscosity, we select the bound of the 2-sided P.I. furthest from the target. A plot similar to Figure 6B can then be constructed using the worst-case values, and then choices about preferred locations in the input space be based on both "expected" and "worst-case" results.

A simulation-based approach that considers the uncertainty more directly is described in the work of Chapman et al,[24] which uses draws from the estimated distribution of the model parameters to characterize the set of all possible response surfaces consistent with the observed data. For each simulation draw, the corresponding Pareto set from Figure 6B is found, and then some graphical summaries from these values are constructed. Figure 7A uses the results from the simulation to examine the robustness of the PF, with larger circles indicating locations appearing more frequently on the front. From this plot, we see that the general shape of the Pareto set remains similar, although a lot of the regions in the input space appear on a PF for at least one simulated set of responses. This should be a reminder to the experimenter that, even though the results from the mean model can be a good indication of general patterns, the uncertainty can change the results considerably.

In the work of the aforementioned author,[24] a multiplicative desirability function was used to combine the scaled responses (based on the best and worst values derived from a 95% prediction interval for the responses) into a single summary. Figure 7B summarizes some results from this subjective phase of the discussion, with both the proportion of times that a design location is selected as best for at least one weight combination for the desirability function (size of square) and the average proportion of desirability function weights for which it is best (shade of gray). The ideal design location would exhibit some robustness with being frequently identified as best and across a diverse set of prioritizations (desirability function weights) of the responses. Figure 7C shows response values for some of the most promising solutions with uncertainty bands. The grid point locations listed along the *x*-axis of the plot correspond to the locations shown in Figure 6B, with larger squares in darker gray. The uncertainty bands were constructed from the 5th and 95th percentiles of values from the simulations at the chosen design location. From this plot, it is possible to see that there are a number of locations for which very similar performance for some of the responses is anticipated. Recall that the goal of the figures in the initial phases of discussion is not to highlight a single solution but rather to identify a manageable number of promising solutions. Using the process described in the work of the aforementioned author,[24] the original 630 grid point locations in the input space have been trimmed to the most promising 21 locations. The graphics in Figures 6 and 7 provide quantitative summaries about their relative strengths and weaknesses across the 3 responses.

Often the decision-making team can agree on some ranges of weights for the different responses. In the case of the chemical process experiment, the team decided to prioritize the responses nearly equally. Figure 8A shows the results for the interested set of weights. The set of dots in the triangle shows which weights were investigated, while the colored

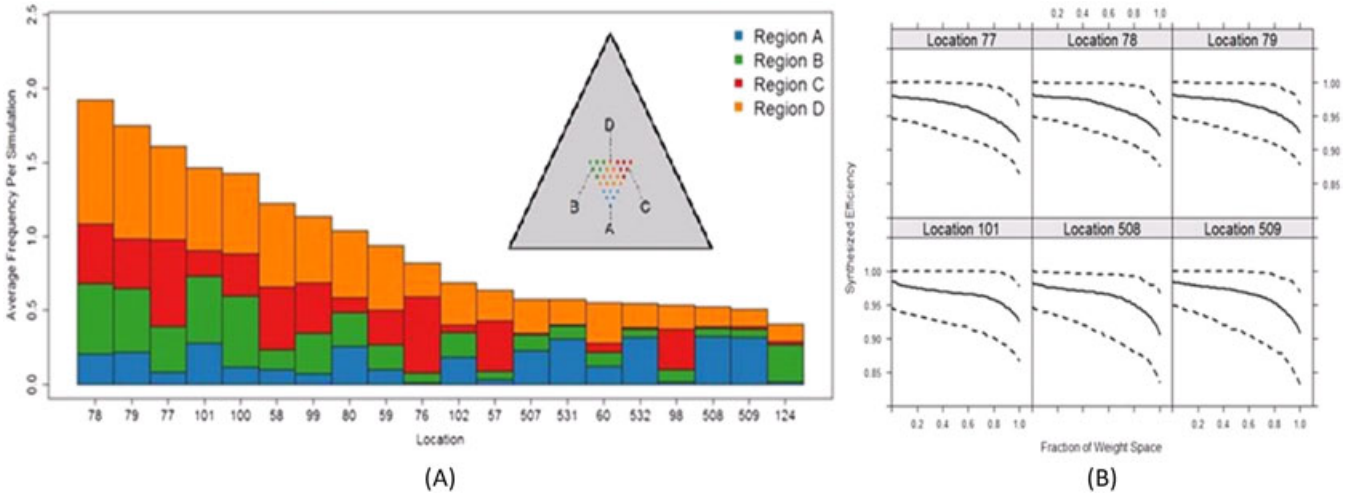

FIGURE 8 A, Mixture plot for showing the average frequency as optimal across the interested weight region for top ranked locations; B, fraction of weight space plot for the lowest synthesized efficiency and its uncertainty for the top solutions [Colour figure can be viewed at wileyonlinelibrary.com]

barplot shows how often different locations are selected as best for different subsets of the chosen weight combinations. This more detailed investigation into performance for a narrow question of interest can help decision makers to pick leading candidates and judge relative performance. Figure 8B shows a version of the FWS plots introduced in Figure 5, but this time, adapted to describe only the subset of preferred desirability function weights and with associated uncertainty. When comparing some of the different locations, the barplot provides more details about how the uncertainty in the response surfaces impacts where best locations are located.

The overall process summarized here and given in more detail in the work of the aforementioned author[24] provide a roadmap to making a structured decision that incorporates the uncertainty in the estimated responses. Including uncertainty achieves 2 important goals. (i) It shows how changes in the response surfaces ripples through to impact the PF and the preferred choices, and (2) it adds realism to what can be expected when a chosen solution is implemented at the conclusion of the experiment.

Some of the plots shown in Figures 6 and 7 will not scale well as the number of responses or the dimension of the input space increases. Some strategies and alternative graphics are suggested in the work of Lu et al[25] to help adapt to more complicated scenarios in higher dimensions. Figure 9 shows a desirability-weight-input-volume (DWIV) plot for

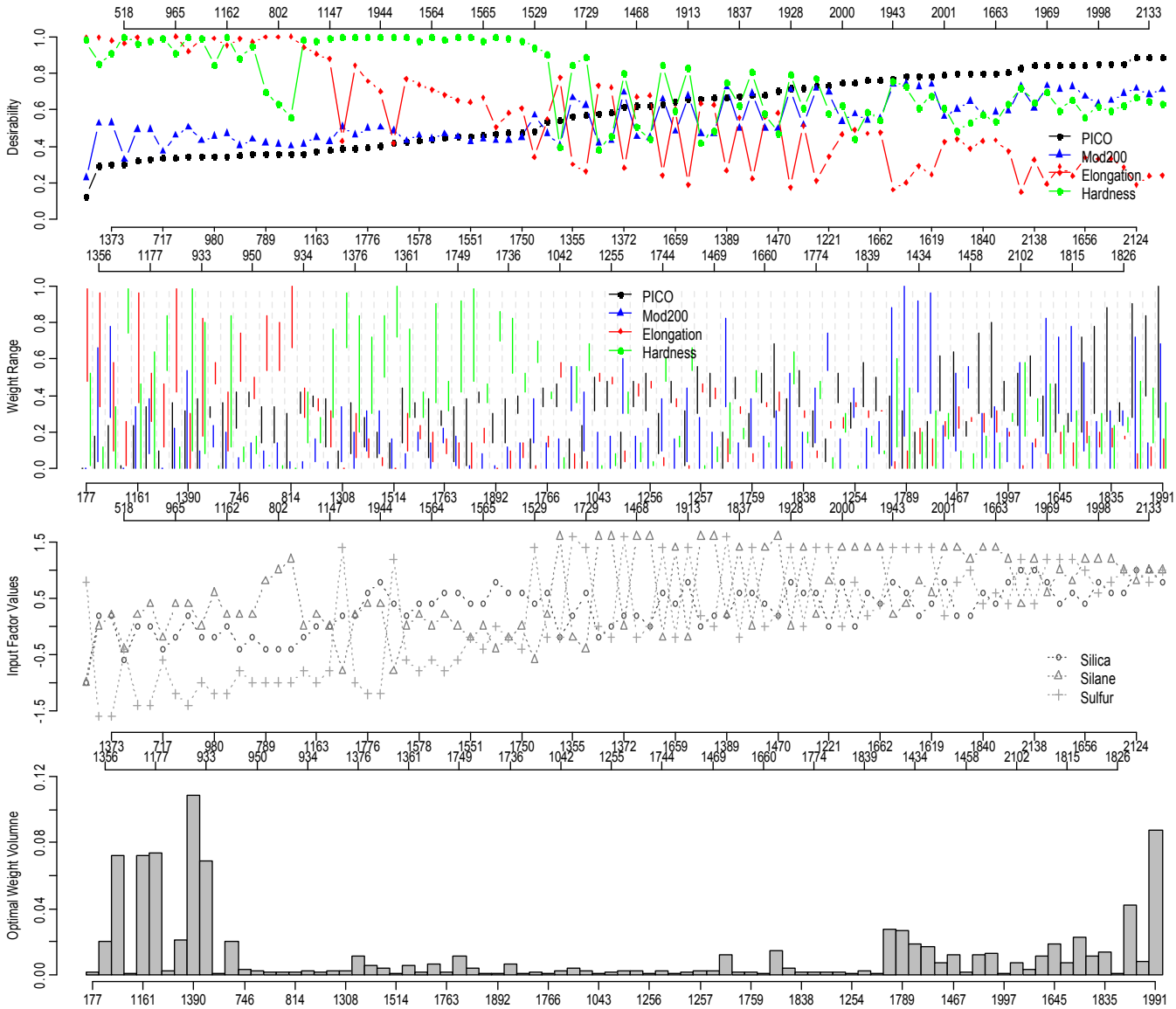

FIGURE 9 The desirability-weight-input-volume plot for the 85 solutions that are optimal for at least 0.5% of all possible weights based on using the multiplicative desirability function and the mean models. The SPs from the top to the bottom show (A), the desirability values of all 4 responses; (B) the range of weights for which a design has optimal desirability value for individual responses; (C) the input factor values, and (D), the fraction of volume of the weights for which a solution is optimal [Colour figure can be viewed at wileyonlinelibrary.com]

4 responses in a 3-dimensional input space problem based off the estimated mean models for each response. The plot is comprised of 4 components, which are joined with a common *x*-axis for the indices of the locations. The first panel (D represents the desirability) shows trade-offs between all of the responses. The second panel (W represents the weights) summarizes robustness of the choices for different desirability function weight combinations (the line segments in different colors indicate the ranges of weights for which a location is optimal for each response). The third panel (I representes the inputs) matches *x*-axis label to the input factor values of a location in the design space, and the last panel (V represents the volume of weights) summarizes the volume of the weight combinations for which a location is best and hence indicates the robustness of the solutions across different weight combinations. While initially a bit overwhelming, this plot allows different patterns to be extracted from the vast grid of candidate input locations. The key is to look at both the bigger patterns across the width of the plot but then to dig into the details of a particular point and understand what information is available for that choice.

A key strategy for using this plot effectively is to use it dynamically to allow different patterns of interested features to be extracted for understanding the trade-offs between responses and how they are connected with the input factor values. The current version of the plot has sorted the *x*-axis based on the desirability for one of the responses, PICO abrasion index (PICO) from worst to best. From this plot, we can extract that the top locations (right of plot) are best for large ranges of weights where PICO is emphasized (from W), that the locations in the input space have all 3 input values between 0.5 and 1.0 (from I), and the top solution for PICO is best for one of the larger proportions of weight combinations (from V). If interest lies in emphasizing a different response, then creating a new version of the plot sorted on the desirability for this response would provide easy access to those insights. Similarly sorting on different *x*-values could reveal patterns in changes in response across the input space. Variations of the DWIV plot can also be adapted to look at subsets of desirability function weights, similar in spirit to Figure 8. The flexibility to explore different versions of the plot dynamically can help to mitigate some of the curse of dimensionality when visualizing is more challenging.

One thing that has been sacrificed with the DWIV plot is the ability to directly investigate the role of uncertainty. This increases the complexity of the problem substantially, and we suggest a sequential approach of doing some preliminary triage of best locations based on the mean model and perhaps a worst case summary and then investigating this smaller set of locations with uncertainty included in the discussion.

We now summarize some of the key points illustrated from this scenario.

G1. Use a process to find common ground. Here, there are many aspects to consider, and following a structured process can help keep the team focused on a common activity and stage of discussion while allowing structured choices to be made throughout.

D2. Build the decision space to include diverse alternatives. In this example, the grid of locations covered the entire design space. By looking broadly initially, it is possible to see if some locations initially not considered important by the mean model might be contenders when uncertainty is included in the decision-making process and vice versa.

R1. Reduce the number of choices on which to do a detailed comparison to a manageable number. The use of the PF allows for a fraction of the choices to be eliminated objectively. Using a desirability function to further streamline the choices based on core priorities that the group can agree upon can also help reduce the number of choices to consider in detail.

C3. Think globally and locally. One of the principles that Tufte[26] encouraged is to reveal data at several levels of detail. Several of the plots included in this section allow for bigger pattern in the data to be seen but then also the detailed comparisons between alternative solutions. Early in the process, the goal is often to identify regions to consider. Later in the process, the goal is to find and be able to justify a particular choice.

C4. Consider dynamic graphics when dimensionality of problem suggests it. The DWIV plot in Figure 8 is a good example of using 1 plot (albeit a relatively complicated one) in many forms to explore and to encourage articulation of team member priorities. Some team members may be more willing to say "let's see what this plot looks like if we prioritize Response X" than they are to say "I think we should prioritize Response X more." However, through the exploration, team members can gain an improved sense of how similar or different their choices are for different priorities.

S1. Create or use a graphical summary that appropriately captures the needed level of detail. In the sequence of graphics shown in this section, many of the graphics were tailored to match a particular goal or query. Having a collection of available graphs from which to choose can help guide discussion and improve the chances for a consensus with a stepwise decision-making sequence. Some of the graphics include some complexity, so during team discussion, it is important to take a few moments when a new graph is introduced to make sure that everyone understands what information is contained in the plot and how to extract it.

S2. Include graphics to formalize conclusions. As the decision-making process draws to a conclusion, it is important to create graphical summaries that allow comparisons between some of the leading contenders. These help to consolidate the team's decision with quantified measure of relative performance. In addition, when the team needs to convince others outside the decision-making process of the merits of the decision, these summaries can be valuable.

6 | PICKING THE TOP *N* CHOICES

Finally, we consider an example where the goal of the decision-making team was not to select a single best choice but rather a set of top *N* choices. We envision 2 common scenarios where this approach might be relevant. (i) Given multiple quantitative criteria, identify the top *N* solutions to accomplish a task; or 2) make a decision that is evaluated based on several primary quantitative criteria and secondary qualitative priorities. An example of the first scenario is described as a case study in the work of Burke et al,[27] while the second scenario might involve choosing a new job. There are a number of quantitative metrics on which to compare the jobs, but there are also some intangible or difficult to quantify measures, which could be used as final tiebreakers, after the leading contenders have been identified.

We now revisit the example described in the work of the aforementioned author,[27] which involved a team of Department of Defense program managers all vying for additional funds for some of their munition stockpiles. The budget allowed for 4 stockpiles to receive these funds. After historical approaches of lobbying for the resources had led to discontentment with the quality of decisions made, an adaptation of the DMRCS process[3] was considered. The team developed a collection of criteria that were considered important to determining which stockpiles were most deserving of receiving these additional funds for surveillance and improvements. After discussing the merits of the different criteria in the Define stage of DMRCS, 3 categories were identified as of primary importance. Historically, the estimated *reliability* curves were the primary focus, and so it was included as a category. A second category was labelled *urgency* and considered the available supply of units relative to the projected demands. Finally, the last category *consequence* considered the impact on the fighting force if an adequate supply of the units was not available.

From these 3 categories, detailed measurable and consistent metrics were developed by subject matter experts (Measure in DMRCS) and data collected across all 42 of the stockpiles under consideration. During the Reduce stage, a layered PF[27] was constructed that identified top contenders that might potentially be in the top 4 choices from among all the stockpiles. This set of fronts eliminated 16 of the stockpiles as not on the top 4 layers of the PF. In the Combine stage, an additive desirability function was selected to combine desirability scores for reliability, urgency, and consequence. Since the panel of experts had scored each of these criteria on a scale of 0 to 10, the team decided to use the natural range of each metric and map the most critical value of 10 to a desirability score of 1, and the least critical value 0 to a core of 0. The team then explored which stockpiles emerged as somewhere in the top 4 most critical across the breadth of possible weightings.

Figure 10 shows a slice of the mixture plot with the top choices for some different weights, ie, w_i. In this figure, Consequence (which is considered the most important among all three categories) is fixed at a weight of 0.4, and the weights for the remaining criteria are varied from 0 to 0.6 (since all 3 weights must sum to 1). From this plot, the top 4 choices can be identified and ranked. For example, for a weighting of $(w_{Rel}, w_{Urg}, w_{Cons}) = (0.25, 0.35, 0.4)$, which is located as the sixth set of points from the left, we can see that stockpile A3 is most critical (darkest shade) followed by B4, A4, and D4 (lightest shade of gray). A copy of the JMP Add-in to identify contending

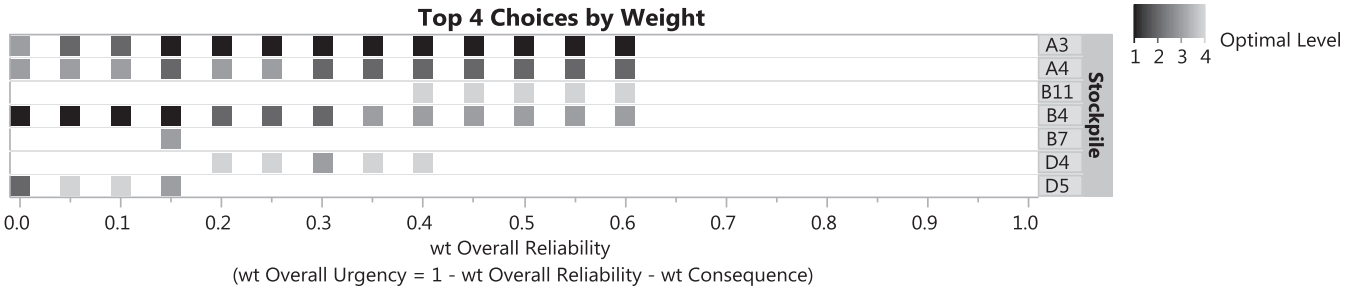

FIGURE 10 Mixture plot for the top 4 choices across different possible weight combinations given a fixed weight for consequence at 0.4 for the stockpile prioritization management example

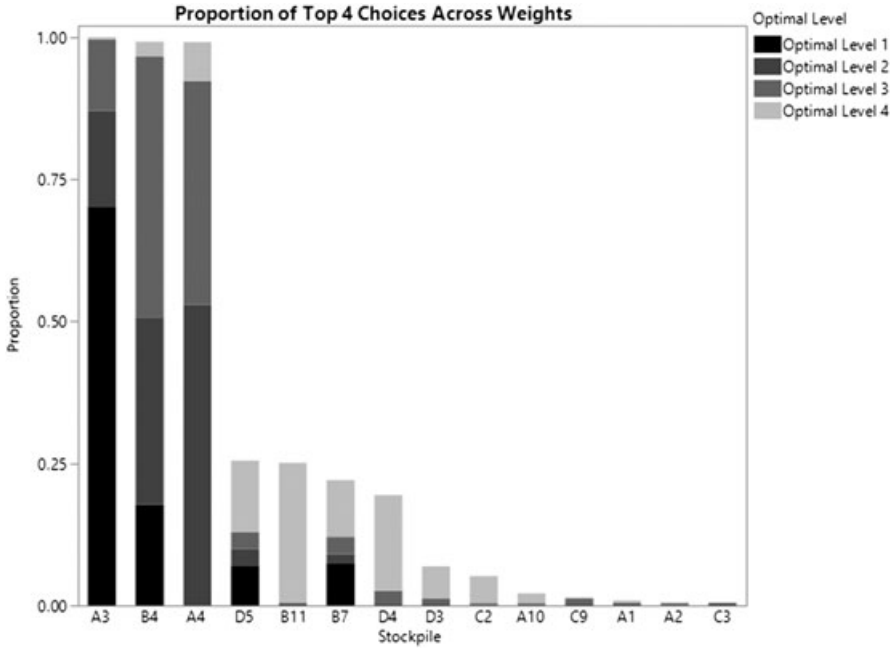

FIGURE 11 Proportion plot summary for top 4 choices across all possible weight combinations for the stockpile prioritization management example

solutions on the top N lays of PFs and create these plots is available at https://community.jmp.com/t5/JMP-Add-Ins/Top-N-Pareto-Front-Search-for-Structured-Decision-Making/ta-p/36527. By examining different weightings of the criteria, it is possible to see how the different stockpiles rank for the priorities of different team members. This facilitated good discussion on the team and allowed members to see how compromise was needed when their emphasis did not agree. It also changed the conversation from individual program managers lobbying for what they wanted to a quantitative discussion of the relative merits of the different choices.

Figure 11 shows the proportion of weights that different stockpiles were ranked in the top 4 choices and how often they were in each position. In the Select stage, the decision-making team needed to identify the 4 stockpiles that were to receive the additional funds. The top 3 choices were quite clear with A3, B4, and A4 being in the top 4 for almost all sets of weights. It is interesting to note that had just a single PF been used, A4 would not have been identified since it was not on the top layer of PF and was never ranked as most critical for any weight choices. However, it clearly belongs in consideration since it is almost always ranked in the top 4 for a broad spectrum of priorities. The final stockpile to be selected was a difficult choice, but the team was able to use a variety of graphics to understand the choices and make a consensus decision. One of the plots that they used was an adaption of Figure 11 but with a more focused subregion of weights, ie, $w_{Rel} \geq 20\%$, $w_{Urg} \geq 10\%$, and $w_{Cons} \geq 30\%$ on which the team had agreed. From this plot, D4 was selected as the fourth stockpile to receive the funds.

Some of the key points illustrated from this scenario include the following.

G1. Use a process to find common ground. In this case, the DMRCS process built consensus by investing early in the process on finding criteria and metrics on which the team could agree. They also include experts to create trustworthy scores for each stockpile. Finally, selecting a subregion of weights on which to focus helped to consolidate the decision.

D1. Define and use an appropriate summary that directly connects to the decision. In this case, 3 categories of criteria were chosen, and metrics that were measurable and comparable across the stockpiles were used. These initial discussions helped define the right scope of the problem and also helped to build consensus about what the decision-making task was all about.

D2. Build the decision space to include diverse alternatives. All of the stockpiles were initially considered, which was helpful to formally eliminate them when they did not appear on the layered PF or ranked in the top 4 for any weight combination. This encourages the team to feel that all stockpiles were duly included and helped to find common areas of agreement as some choices were eliminated.

C1. Allow common visualization and discussion about results, instead of keeping subjective elements of the decision unshared. For the example, the mixture and proportion plots in Figures 10 and 11 allowed the team to identify key stockpiles to focus on and judge the impact of their subjective choices. This turned what had been previously a negotiation into a data-driven quantitative comparison and discussion.

S2. Include graphics to formalize conclusions. When the team concluded their decision-making process, the team members had summary plots to share with others not involved in the process. These provided compelling evidence of how the data-driven evaluation process was conducted and how the decisions were reached.

7 | CONCLUSIONS AND DISCUSSION

Statistical thinking[5] and statistical engineering[28-30] emphasize how good process and including data with its associated uncertainty are important for solid business and industry practice. Carefully designed graphics based on metrics important to the practitioner can be extremely helpful for serving these purposes. Much of statistical engineering has focused on problem solving for product and process improvement, but there are other aspects of problem solving for which strategies, tactics, and tools are needed. Introducing data-driven quantitative summaries into discussions and the decision-making process and presenting them in efficient and informative graphics to aid understanding and discussion is key to good outcomes.

Most decisions have a data component and a subjective component. What do we know? Moreover, how should we balance the trade-offs when optimal for all facets of the decision are not simultaneously achievable? When making decisions as an individual, we have developed personalized methods for assessing alternatives, balancing our priorities, and often use informal or qualitative summaries of the choices. To make team decision making work and avoid unsatisfactory decisions driven by personalities instead of data and facts, there needs to be added structure to the process that is built around strategies and tools that promote discussion and build fundamental consensus about the goals of the decision. To effectively explore alternatives, there need to be ways to allow comparison between options, encourage sharing of individual priorities, and summaries that provide insight about how robust choices are to these different prioritizations. Good graphical summaries can provide an effective platform to extract information and support understanding, a mechanism to balance the contributions of different facets of a decision, and a forum for discussion and reaching consensus.

In this paper, we outlined 15 principles that we think can improve the focus of discussion and decision making, which are summarized in Table 1. At first glance, many of the principles may not seem directly graphics focused. However, they are actually closely relevant to how we build and use the graphical summaries to support improving decision making. It is our hope that providing more detailed examples of how the principles apply in practice will help them to become more concrete. We recommend using this list as the guiding principles for designing graphical tools to facilitate discussion and informed decision making.

REFERENCES

1. Anderson-Cook CM, Graves T, Hamada M, et al. Bayesian stockpile reliability methodology for complex systems. *J Mil Oper Res Soc*. 2007;12(2):25-37.
2. Anderson-Cook CM, Graves T, Hengartner N, et al. Reliability modeling using both system test and quality assurance data. *J Mil Oper Res Soc*. 2008;13:5-18.
3. Anderson-Cook CM, Lu L. Much-needed structure: a new 5-step decision-making process helps you evaluate, balance competing objectives. *Qual Prog*. 2015;48(10):42-50.
4. Anderson-Cook CM. Optimizing in a complex world: statisticians' roles in decision-making (with discussion and rejoinder). *Qual Eng*. 2017;29(1):27-41.
5. Hoerl R, Snee R. *Statistical Thinking: Improving Business Performance*. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2012.
6. Lu L, Anderson-Cook CM, Robinson TJ. Optimization of designed experiments based on multiple criteria utilizing a Pareto frontier. *Technometrics*. 2011;53:353-365.
7. Hamada MS, Wilson AG, Reese CS, Martz HF. *Bayesian Reliability*. New York, NY. Springer; 2008.
8. Lu L, Anderson-Cook CM. Prediction of reliability of an arbitrary system from a finite population. *Quality Eng*. 2011;23:71-83.
9. Anderson-Cook CM, Lu L. Best bang for the buck-part 1: the size of experiments relative to design performance. *Qual Prog*. 2016;49(10):45-48.
10. Anderson-Cook CM, Lu L. Best bang for the buck-part 2: choosing between different sized designs. *Qual Prog*. 2016;49(11):50-52.
11. Jones B, Nachtsheim CJ. A class of three-level designs for definitive screening in the presence of second-order effects. *J Qual Technol*. 2011;43(1):1-15.

12. Lenth RV. Some practical guidelines for effective sample size determination. *Am Stat*. 2001;55(3):187-193.
13. Zahran A, Anderson-Cook CM, Myers RH. Fraction of design space to assess prediction capability of response surface designs. *J Qual Technol*. 2003;35(4):377-386.
14. SAS Institute Inc. JMP, version 13. 2016.
15. 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):404-422.
16. 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 Stochastic Models Bus Ind*. 2012;28(3):206-221.
17. Goos P. *The Optimal Design of Blocked and Split-Plot Experiments*. New York, NY: Springer; 2002.
18. 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.
19. Derringer G, Suich R. Simultaneous optimization of several response variables. *J Qual Technol*. 1980;12(4):214-219.
20. 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. *Technometrics*. 2013;55(4):473-487.
21. Lu L, Anderson-Cook CM, Lin D. Optimal designed experiments using a Pareto front search for focused preference of multiple objectives. *Comput Stat Data Anal*. 2014;71:1178-1192.
22. Chapman JL, Lu L, Anderson-Cook CM. Process optimization for multiple responses utilizing the Pareto front approach. *Qual Eng*. 2014;26(3):253-268.
23. Myers RH, Montgomery DC, Anderson-Cook CM. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. New York, NY: Wiley; 2016.
24. Chapman JL, Lu L, Anderson-Cook CM. Incorporating response variability and estimation uncertainty into Pareto front optimization. *Comput Ind Eng*. 2015;76:253-267.
25. Lu L, Chapman JL, Anderson-Cook CM. Multiple response optimization for higher dimensions in factors and responses. *Qual Reliab Eng Int*. 2016;33(4):727-744.
26. Tufte ER. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press; 2001.
27. Burke SE, Anderson-Cook CM, Lu L, Montgomery DC. Prioritization of stockpile maintenance with layered Pareto fronts. *Qual Eng*. 2018. https://doi.org/10.1080/08982112.2017.1390585
28. Hoerl RW, Snee RD. Closing the gap: statistical engineering links statistical thinking, methods, tools. *Qual Prog*. 2010;43(5):52-53.
29. Anderson-Cook CM, Lu L, Clark G, et al. Statistical engineering – forming the foundations. *Qual Eng*. 2012;24(2):110-132.
30. Anderson-Cook CM, Lu L, Clark G, et al. Statistical engineering – roles for statisticians and the path forward. *Qual Eng*. 2012;24(2):133-152.

How to cite this article: Anderson-Cook CM, Lu L. Graphics to facilitate informative discussion and team decision making. *Appl Stochastic Models Bus Ind*. 2018;34:963–980. https://doi.org/10.1002/asmb.2325