**Optimal Experimental Designs for Hypothesis Testing with Multiple Factors**

Colin M. Lynch1, Douglas C. Montgomery2

1School of Life Science, Arizona State University, Tempe, AZ, USA 85287

2School of Computing and Augmented Intelligence, Arizona State University, Tempe, Arizona, USA 85287

**Abstract**

The power of an experiment is often only calculated for a single factor given assumptions on effect size, sample size, and the significance level. However, when an experiment includes multiple factors, then the spread of points among those factors could influence power. Here, we use JMP to calculate the power of different designs when an experimenter requires a detailed mapping of an experiment’s response surface. We also consider the case where the experimenter only wants to measure the significance of the main effects of different predictor variables. In the former context, we find that A-efficient designs tend to maximize power. Central composite designs in particular provide the highest power levels. In the latter context, we find that different first-order designs provide comparable levels of power; however, they do so for differing numbers of factors. Definitive screening designs maintain high power levels for the largest number of factors.

**Keywords**

Power, experimental designs, hypothesis testing, classical designs, screening experiments, response surface, multiple factors, continuous variables, optimal design, A-Efficiency, sample size

**Bibliographic notes:** Colin M. Lynch is an Animal Behavior Ph. D. candidate at Arizona State University. He is an NSF GRFP Fellow and received his bachelor’s degree in Neuroscience from the University of Arizona. He is broadly interested in optimizing sampling techniques in the context of complex adaptive systems.

Douglas C. Montgomery is a Regents’ Professor and Foundation Professor of Engineering at Arizona State University. He is a Fellow of the ASQ, the ASA, the RSS, the IIE, a member of the ISI, and an Academician of the IAQ. He received the Deming Lecture Award from the ASA. the Shewhart Medal, the William G. Hunter Award, the Brumbaugh Award, the Lloyd S. Nelson Award, Shewell Award (twice) from the ASQ, the George Box Medal from ENBIS, and the Ellis Ott Award. He is the Chief Editor of *Quality and Reliability Engineering International*.

1. **Introduction**

Arguably, one of the primary goals of any experiment is to test hypotheses. That is, experiments help scientists determine whether or not a factor has a significant effect on a response variable of interest. However, any experimental signature is likely to be affected by numerous factors, and these factors in many cases interact (Fraker and Peacor, 2008; Diester et al., 2019). Despite this, many experiments focus on only one or two factors at a time, and tools designed to aid in experimental design (such as the power analyses used in G\*Power; Kang, 2021) only focus on single effects when determining sample size. While there are strategies for testing multiple hypotheses simultaneously (i.e. multiple comparisons), these strategies usually focus on assessing the differences in a response variable across different levels of the same factor (Michel et al., 2020). When there are multiple factors, information criteria (such as AIC) can be used to evaluate models that test multiple hypotheses simultaneously (Dochtermann & Jenkins, 2011), however, these methods say nothing about how to set factor levels in the first place. Therefore, a general framework for selecting design points across multiple factors for hypnosis testing is lacking.

Fortunately, such a framework exists for other problem sets. Experiments do not only need to test hypotheses, but they could establish standard operating conditions, maximize a response variable, and increase the robustness of the signal (Myers et al., 2016). Such objectives require minimizing (or otherwise manipulating) scaled prediction variation (SPV), which describes the error associated with making a prediction in a regression model. SPV is scaled by the variation of the response and is multiplied by the sample size so that it can be used as a metric to compare designs directly (Montgomery, 2020). Different optimization criteria could be used to minimize different aspects of SPV. For instance, the G-optimality criterion minimizes the maximum prediction variance in the region of an experimental design (Kiefer, 1959). Conversely, other optimality criteria, such as A-optimality, minimize the sum of variances of the regression coefficients (Kiefer, 1959). This latter set of criteria could be useful for determining which factors have a significant effect on the response variable, as lowering the variation in coefficients inherent to the design could increase the signal-to-noise ratio of corresponding F-tests or 𝛘2 tests. This could increase the probability of correctly rejecting the null hypothesis, thereby implicitly testing the alternative hypothesis (Cohen, 1992). Therefore, designs that maximize A (or D) efficiency couldbe the most suitable for scientists who want to test multiple hypotheses.

For instance, many factors can influence bone density in animals such as diet, lifestyle choices, and age (Sanders et al., 2014). These factors can operate independently or through interactions with one another. For instance, calcium cannot be properly absorbed into the intercellular matrix without vitamin D (Goltzman et al., 2018) and the effects of exercising with calcium supplements can be multiplicative rather than just additive (Specker, 1996). Studies that focus on such complicated signatures could benefit greatly from multiple-factor designs so that meaningful interactions are not missed. As many of these studies require also longitudinal experiments where factor levels are set by raising mouse models in the lab, each individual measurement is expensive. Therefore, these types of studies could also benefit from optimal designs which seek to minimize sample size while keeping power high.

In this paper, we study the ability of different experimental designs to maximize power in two different contexts. In the first context, we assume that scientists are interested in detecting interactions among continuous factors as well as higher-order effects, and will therefore fit a second-order model. We perform power analyses on classical and computer-generated designs which can accommodate a second-order regression model, and we measure the efficiencies of each design to determine their effect on power. In the second context, we assume that scientists are only interested in detecting the main effects of different continuous variables and only fit a first-order model. Screening experiments are ideal for such cases, so we study the effect of each screening design on power.

1. **Methods**

To measure the relationship between power and optimality for second-order designs, we perform power analyses on classical designs (Box-Behnken, rotatable central composite design, orthogonal central composite design) as well as computer-generated designs (A, D, I, and alias optimal) across sample sizes, effect sizes, and factor numbers. We calculate the D, A, and G-efficiencies of each design and we measure the power associated with the intercept (𝛃0), a randomly-selected main effect (𝛃i), a randomly-selected interaction (𝛃ij), and a randomly-selected second order term (𝛃i2). Sample sizes were determined in two different ways. After a factor number is selected, we 1) measure power after choosing the recommended sample size for that experiment (as determined by the program JMP), and 2) we match the sample sizes across experiments within a particular factor number. This usually includes adding replicates or center runs to a design. We ran a full factorial for each sample size method, calculating power and efficiency across all combinations factor numbers, k ∈ {2, 3, 4, 5}, each type of coefficient (𝛃0, 𝛃i, 𝛃ij, 𝛃i2), each effect size 𝛃 ∈ {0, 0.25, 0.5, …, 2}, and each type of experimental design (Alias, A, I, D optimal, Box-Behnken, orthogonal and rotatable central composite design). We therefore evaluated 4\*4\*9\*7 = 1,008 designs.

In the first-order model case, we compare the effectiveness of different screening designs (2k, 2k-p, definitive screening), as these are all either near-optimal or optimal designs (each design has a D, A, and G efficiency between 90% and 100% for a first-order model). As these optimality criteria are extremely similar, we do not compare the relative efficiencies of each design. We only compare the powers of different screening experiments across different sample sizes and for differing numbers of factors. Power was calculated for the intercept (𝛃0) and a randomly-selected main effect (𝛃i).

Sample sizes for these designs are less flexible than their second-order counterparts, so to compare power on a per-run basis we compared all designs within a run limit (n ∈ {4, 5, …, 32}). These designs included 2k where k ∈ {2, 3, 4, 5}, definitive screening designs where k ∈ {5, 7, 9, 11}, and fractional factorials 23-1 (resolution III), 24-1 (resolution IV), 25-1 (resolution V), 25-2 (resolution III), 26-2 (resolution IV), 26-1 (resolution V), 26-3 (resolution III), 27-3 (resolution IV), and 27-2 (resolution IV). It should be noted that resolution III designs confound main effects with two-factor interactions, but here we assume that the main effects are more likely to occur than the interactions. Power was calculated across each type of coefficient (𝛃0, 𝛃i) and each effect size 𝛃 ∈ {0, 0.25, 0.5, …, 2}, so there was a total sample size of 2\*9\*(4+4+9) = 306.

Designs were generated in JMP Pro 16.2 (SAS Institute Inc., Cary, NC, USA), and the power, D, A, and G efficiencies of these designs were also calculated in JMP. Analyses of these measurements were performed in R studio (R Core Team, 2021) using the base stats and FSA packages while ggplot2 and ggpubr were used for graphical displays.

1. **Power in second-order designs**

We first want to know which type of efficiency is the most correlated with power. D, A, and G efficiencies are all highly collinear (VIF = 7.257, 3.682, 3.568, respectively), so we cannot build a linear model with all 3 as independent variables. Instead, we build 3 separate models with each type of efficiency as a predictor variable. In each of these models, power is the response variable while effect size, sample size, and coefficient type (𝛃0 vs 𝛃i, etc.) are the other predictor variables. We only considered the main effects of these predictor variables. Factor number was insignificant in each of these models (LM: t ≅ -0.6, p-value > 0.05), so it was not included.

Each of these models were highly significant (D-Efficiency LM: F6, 1504 = 996, p-value < 0.001; G-Efficiency LM: F6, 1504 = 992.2, p-value < 0.001; A-Efficiency LM: F6, 1504 = 1021, p-value < 0.001). Each efficiency was also positively correlated with power (D-Efficiency LM t = 5.7, G-Efficiency t = 5.268, A-Efficiency t = 7.981). Among these models though, A-Efficiency performed the best, followed by D and then G-Efficiency (D-Efficiency LM R2 = 0.798, PRESS = 43.858, G-Efficiency LM R2 = 0.797, PRESS = 43.982, A-Efficiency LM R2 = 0.802, PRESS = 42.99). A-Efficiency also has the lowest AIC value, indicating that it has the best fit (D-Efficiency AIC = -1059.071, G-Efficiency AIC = -1054.404, A-Efficiency AIC = -1089.451).

We next want to determine which designs yield the highest power, and whether or not these designs also had the highest A-Efficiency. To do this, we first performed an ANOVA between designs and power (F6, 1805 = 2.188, p-value < 0.05). We then performed a post-hoc Tukey test to determine which designs had the highest power (significance groups are given in Fig. 1). Generally, we find that the central composite designs had the highest power.

Next, we ranked each design based on how high its average power was as well as how high its average efficiency (D, A, G) score was (in all cases, high mean across runs = higher rank). We then perform a Spearman’s rank correlation test between power rank against each efficiency rank (Fig. 2). Only the correlation between A efficiency and power is significant (Spearman’s Rank correlation: A-efficiency rho = 0.271, p-value < 0.05; D-efficiency rho = 0.255, p-value = 0.057; G-efficiency rho = 0.131, p-value = 0.337).

1. **Power in first-order designs**

We are primarily interested in determining whether or not different screening designs yield higher powers for the main effects and intercept of a model at different sample sizes. To do this, we performed a LOESS regression between the number of runs and power for each type of design (Fig. 3A). We calculated the 95% confidence intervals of the regression, and found that these intervals overlap across all sample sizes and across designs. This indicates that on a per-run basis, none of these designs result in a higher power than another. Despite this overlap, some designs provide the same power for more factors than others (Fig. 3B). Fractional factorial designs can screen more factors than full factorial designs, and when there are more than 20 runs, definitive screening designs can test for the highest number of factors.

1. **Concluding remarks**

When an experiment includes multiple factors, power should not be calculated on the basis of a single factor, as the spread of the design points will influence the power levels of all factors. We found that design optimality can be co-opted as a framework for finding designs that maximize power. Specifically, A-efficiency was positively associated with power for second-order models. Central composite designs (both orthogonal and, to a slightly lesser extent, rotatable) had the highest power while also having the highest A-efficiency, so if nothing about the model system is known *a priori,* these designs should be chosen for hypothesis testing if all factors are continuous.

If an experimenter is only interested in detecting the main effects of an experiment, then screening designs should be used instead, as similar levels of power can be achieved at smaller sample sizes. Each of the screening designs tested here yielded similar power on a per-run basis, however, definitive screening designs retain this power for more factors. If only a couple of factors are present in an experiment, a full factorial design should be used, as it has a number of desirable properties that other designs do not have (effects are not aliased, the design is rotatable, etc.). If a moderate number of factors are present, then fractional factorials should be used instead. Finally, if there are a large number of factors, then definitive screening is the best option.

**References**

Cohen, J. (2016). ‘A power primer.’ *Psychological Bulletin,* Vol. 112 No. 1, pp. 155–159. DOI: 10.1037//0033-2909.112.1.155

Diester, C. M., Banks, M. L., Neigh, G. N., & Negus, S. S. (2019). ‘Experimental design and analysis for consideration of sex as a biological variable.’ *Neuropsychopharmacology*, Vol. 44 No. 13, pp. 2159-2162. <https://doi.org/10.1038/s41386-019-0458-9>

Dochtermann, N. A., & Jenkins, S. H. (2011). ‘Developing multiple hypotheses in behavioral ecology.’ *Behavioral Ecology and Sociobiology*, Vol 65 No. 1, pp. 37-45. https://doi.org/10.1007/s00265-010-1039-4

Fraker, M. E., & Peacor, S. D. (2008). ‘Statistical tests for biological interactions: a comparison of permutation tests and analysis of variance.’ *Acta Oecologica*, Vol. 33 No.1, pp. 66-72. https://doi.org/10.1016/j.actao.2007.09.001

Kang, H. (2021) ‘Sample size determination and power analysis using the G\* Power software.’ *Journal of educational evaluation for health professions,* Vol. 18. DOI: https://doi.org/10.3352/jeehp.2021.18.17

Kiefer, J. (1959). ‘Optimum Experimental Designs.’ *Journal of the Royal Statistical Society, Series B* Vol. 21 No. 2, pp. 272-319. https://doi.org/10.1111/j.2517-6161.1959.tb00338.x

Michel, M. C., Murphy, T. J., & Motulsky, H. J. (2020). ‘New author guidelines for displaying data and reporting data analysis and statistical methods in experimental biology.’ *Journal of Pharmacology and Experimental Therapeutics*, Vol. 372 No. 1, pp. 136-147. DOI: https://doi.org/10.1124/jpet.119.264143

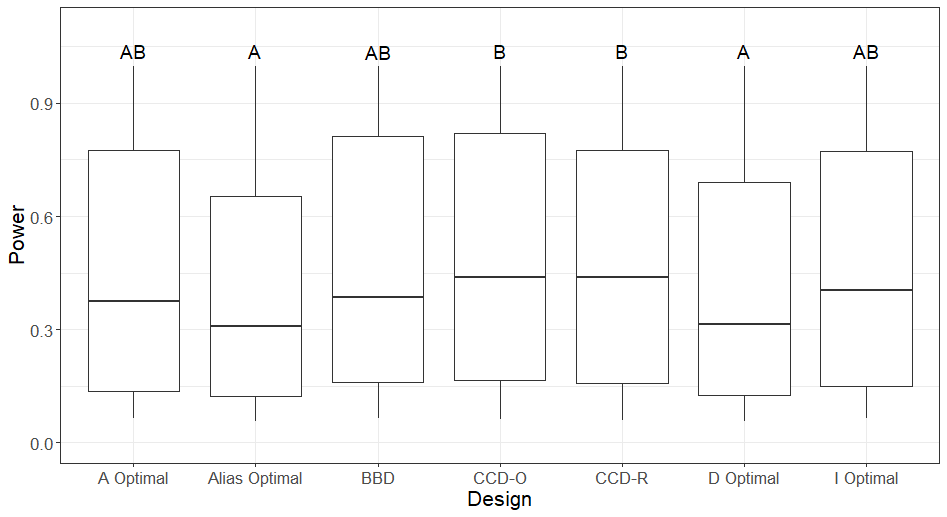
Montgomery, D. C. (2020). *Design and analysis of experiments*. 10th edition, John wiley & sons, Hoboken, NJ.

Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2016). *Response surface methodology: process and product optimization using designed experiments.* 4th edition, John Wiley & Sons, Hoboken, NJ.

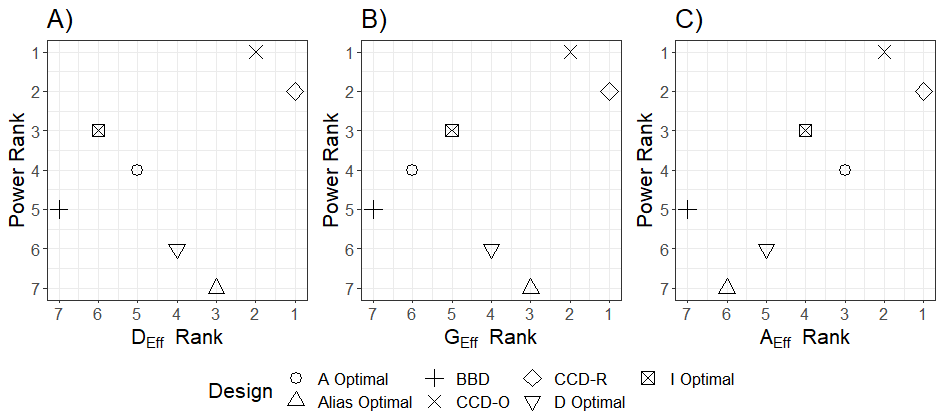
R Core Team (2021). ‘R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.’ URL [http://www.R-project.org/](http://www.r-project.org/)

Sanders, Kerrie M., David Scott, and Peter R. Ebeling. (2014). ‘Vitamin D deficiency and its role in muscle-bone interactions in the elderly.’ *Current osteoporosis reports*, Vol. 12 No.1, pp. 74-81. https://doi.org/10.1007/s11914-014-0193-4

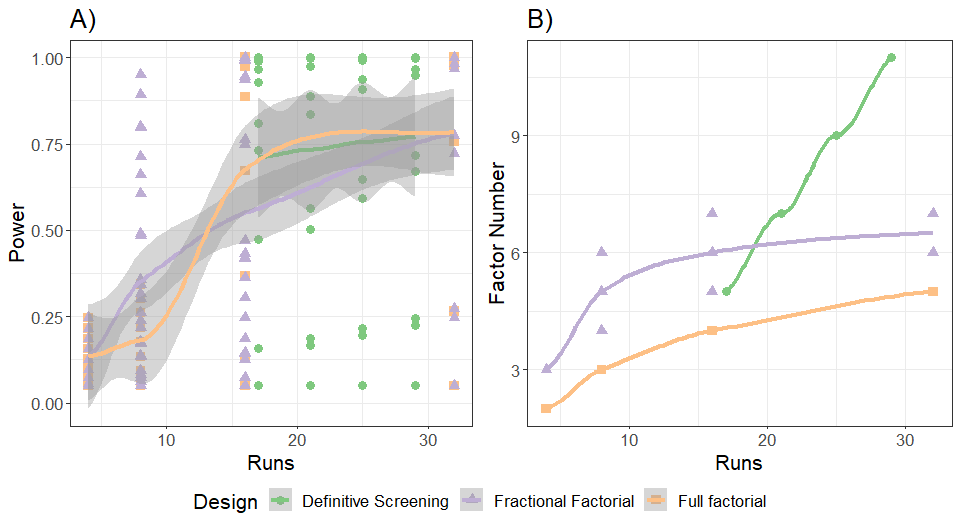
Specker, B. L. (1996). ‘Evidence for an interaction between calcium intake and physical activity on changes in bone mineral density.’ *Journal of Bone and Mineral Research*, Vol. 11 No. 10, pp. 1539-1544. https://doi.org/10.1002/jbmr.5650111022



**Figure 1** Power across all sample sizes, factor numbers, effect sizes, and coefficients (𝛃0, 𝛃i, etc.) for each design (BBD = Box-Behnken design, CCD = central composite design, O = orthogonal, R = rotable). Letters signify significance groups across designs. Group A had the lowest power, group B had the highest, and group AB was indistinguishable from either A or B (Tukey test p-value > 0.05). Whiskers in the plot extend from 0.056 to 0.999.



**Figure 2** Correlation between power rank and A) D-efficiency rank, B) G-efficiency rank, and C) A-efficiency rank for different designs (shape). BBD = Box-Behnken Design, CCD = central composite design, R = rotatable, O = orthogonal.



**Figure 3**  A) LOESS regressions for each design (shape and color) between run number (x-axis) and power (y-axis). The solid line gives the regression line while the shaded region shows the 95% confidence interval for each factor. B) LOESS regressions for each design between run number (x-axis) and the number of factors (y-axis). As factor number did not vary within runs for full factorial or definitive screening designs, confidence intervals could not be calculated.