*Proceedings of the 2001 Winter Simulation Conference*
*B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, eds.*

# GRAPHICAL METHODS FOR ROBUST DESIGN OF A SEMICONDUCTOR BURN-IN PROCESS

Scott L. Rosen
Chad A. Geist
Daniel A. Finke
Jyotirmaya Nanda
Russell R. Barton

310 Leonhard Building
The Harold & Inge Marcus
Department of Industrial and Manufacturing Engineering
The Pennsylvania State University
University Park, PA 16802, U.S.A.

## ABSTRACT

Discrete-event simulation is a common tool for the analysis of semiconductor manufacturing systems. With the aid of a simulation model, and in conjunction with sensitivity analysis and metamodeling techniques, robust design can be performed to optimize a system. Robust design problems often include integer decision variables. This paper shows a graphical approach to robust design that is effective in the presence of discrete or qualitative variables. The graphical robust design methodology was applied to a backend semiconductor manufacturing process. Changes in specific resource capacities and product mix were examined to determine their effect on the level and variance of cycle time and work in process.

## 1 INTRODUCTION

The manufacturing of semiconductor products consists of four distinct stages: wafer fabrication, wafer sort and test, packaging and assembly, and functional (electrical) tests. These four stages can be divided into two categories. Wafer fabrication and wafer sort and test are referred to as front-end manufacturing operations. Packaging, assembly, and functional testing are referred to as backend manufacturing operations.

Infineon is interested in system evaluations on a semiconductor backend manufacturing plant in Malacca, Malaysia. Due to the complexity of the process and high level of randomness of the process, system evaluations on their process can only be accurately done using simulation. Therefore, they have constructed a highly detailed simulation model of their backend process using Factory Explorer. Their simulation model consists of three products, thirteen tool groups, zero operator groups, and three processes. The three products include: TSOP 50, a 0.25 micron process 16 megabit synchronous DRAM; the TSOP 54, a 0.20 micron process 64 megabit synchronous DRAM; and the TSOP 256, a 0.20 micron process 256 megabit synchronous DRAM.

The simulation model analyzes two functions of the backend manufacturing facility, the burn-in process and the test process. The following three steps summarize the burn-in process. First, pre-packaged product from the assembly area is loaded on burn-in boards unique to each product. Then the chips are tested under extreme temperature conditions in batch-process ovens having no special dedication requirements. Finally, the chips are automatically unloaded from the burn-in boards. The loading and unloading steps are performed by the same equipment, resulting in a reentrant flow situation where parts waiting to be processed and parts having already completed processing compete for resources.

Upon completion of the burn-in process, each chip must proceed to the functional test area to be tested. Due to the high cost of failure of semi-conductors, chips are put through rigorous performance tests that simulate real operating conditions. To perform each of these tests a dedicated combination of a tester and two handlers is used. The chips are tested at two different temperature extremes, -10 and 85 degrees Centigrade. In general, the cold test is performed first, followed by the hot test—this phenomenon is due to the high changeover costs, in terms of setup time, because both tests are performed on the same pieces of equipment. A process flow diagram is provided in Figure 1.

Through discussions with Infineon representatives, it was determined that the simulation responses of average cycle time and average work in process (WIP) were the

most critical. More specifically, average cycle time and average work in process pertaining to an average over all three part types of their process: TSOP50, TSOP54, and TSOP 256. These averages were calculated over a period one-week. Ten decision variables of the manufacturing process were believed to have a possible effect on the two model responses of concern. Infineon wanted to examine the effect of capital equipment expenditures on these responses. For example, what would be the effect of purchasing one extra oven on the decrease in cycle time? Also of concern was the effect of product mix. The TSOP50 part is currently being phased out of the process. So, would an increase in flow of the TSOP54 part and a decrease in flow of the TSOP50 part have any effect on work in process or cycle time? Another possible factor affecting cycle time and work in process was the percent of hot lots in the system.

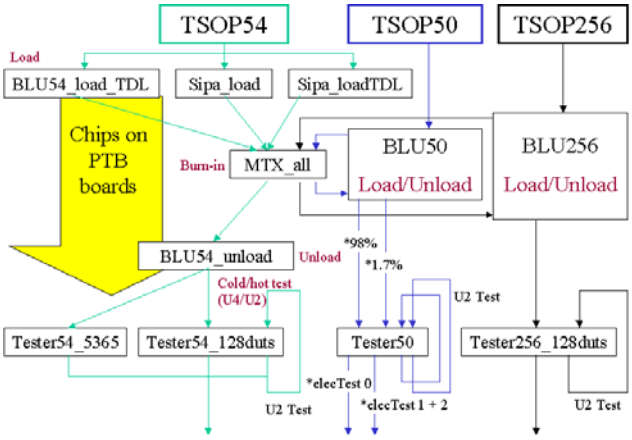


Figure 1: Process Flow

This paper focuses on the implementation of robust parameter design to the Infineon backend semiconductor manufacturing process. With the use of the burn-in simulation model, cycle time and work in process objectives are analyzed by sensitivity analysis and metamodeling. A screening process is performed on the possible factors of influence to remove the factors of insignificance thereby facilitating the robust design process. The fewer variables involved in the robust design process, the less experimental time and computational time is needed. The remainder of the paper is outlined as follows. First, we will discuss the sensitivity analysis, which was used to determine the significance of the factors on each of the two performance measures of interest. The next section contains an introduction to robust parameter design, followed by the results obtained from our backend semiconductor case.

## 2 SENSITIVITY ANALYSIS AND SCREENING

A global sensitivity analysis was performed to determine the possible influence that each factor had on the two responses: work in process and cycle time. Littles's Law relates work in process and cycle time; therefore, only WIP is considered for the remainder of this paper. The analysis involved the construction of first order linear model for work in process and a first order linear model for cycle time from the ten factors. Then statistical tests for the level of significance for each of the ten factors pertaining to each model can be performed

The first step of this analysis involves a definition of the design range for each of the ten variables. Since there is such a large quantity of burn-in boards, the increase in burn-in boards was represented as a percent with the upper limit set to 10% and the lower limit set to 0%. For all of the other variables representing an increase in equipment capacity the upper limit was set at 1 and the lower limit was set at 0. The rationale behind this was based on the assumption that Infineon would not choose to purchase more than one piece of any type of machine. The lower and upper limits of percent hot lots were set at 5% and 30% respectively because the percent of hot lots is not believed to be significant until at least 30% are in the system. Also, the limits for the percent increase and decrease in product mix were set at 2% and 9% because the model became unstable when less than 2% TSOP50s were in the process flow. Accordingly, 2% was set at the lower limit and 9% was set at the upper limit (9% corresponds to the current percentage of TSOP50's in the system). All of the variables were scaled by assigning a "-1" to the lower limit and a "1" to the upper limit. Table 1 specifies each decision variable as well as the respective upper and lower limits for the design variables.

Table 1: Sensitivity Analysis Variables and Ranges of Interest

| Variable Notation | |
|---|---|
| A = %Increase of Burn-in Boards | (-1,1) = (0%, 10%) |
| B = Increase in TSOP54 Loaders | (-1, 1) = (0, 1) |
| C = Increase in TSOP256 Loaders | (-1, 1) = (0, 1) |
| D = Increase in TSOP54 Unloaders | (-1, 1) = (0, 1) |
| E = Increase in TSOP256 Unloaders | (-1, 1) = (0, 1) |
| F = Increase in MTX Ovens | (-1, 1) = (0, 1) |
| G = Increase in TSOP54 Testers | (-1, 1) = (0, 1) |
| H = Increase in TSOP256 Testers | (-1, 1) = (0, 1) |
| J = %Hot Lots | (-1, 1) = (5%, 30%) |
| K = Product Mix | (-1, 1) = (2%, 9%) |

The next step of the sensitivity analysis involves the selection of an experimental design that will provide sufficient information about the decision variables to build the first order regression model. Each simulation run required 15 minutes, making it expensive to run. So the most important criteria in finding an appropriate design was minimizing simulation runs while still obtaining a measure of the significance of each of the main effects to the response. This would require a design with at least N = k + 1 runs

where k is the number of factors, and N is the number of runs needed. Due to the familiarity of fractional factorial designs, a 1/64 fractional factorial was used. This requires 16 runs given the inclusion of 10 factors. This is a resolution III type design meaning that none of the main effects were aliased with each other, however all of the main effects were aliased with two-way interactions. Figure 2 is a graphical representation of the 1/64 factorial design used for Sensitivity Analysis (Barton 1999).

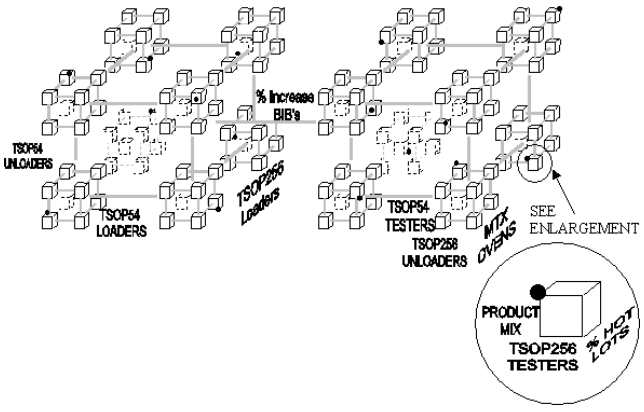


Figure 2: Graphical Representation of a $2^{10-6}$ Fractional Factorial Design

The first order model of work in process is shown in the equation below. This model contains scaled decision parameters.

$$WIP = 1{,}224{,}378 - 296{,}013\ A - 298{,}730\ B - 7{,}244\ C + 4{,}536\ D - 6{,}842\ E - 514{,}976\ F - 532{,}049\ G + - 225\ H\ 300{,}377\ J + 542{,}896\ K \quad (1)$$

Table 2 shows the results of the regression analysis. A test of the significance of regression was also performed on this model and a p-value of .029 was obtained, which shows that overall the model is a good fit (Devore 1995).

Table 2: Regression of WIP vs. A, B, C, D, E, F, G, H, J, K

| Pred | Coef | SE Coef | T | P |
|---|---|---|---|---|
| Int. | 1,224,378 | 134,221 | 9.12 | 0.000 |
| A | -296,013 | 134,221 | -2.21 | 0.079 |
| B | -298,730 | 134,221 | -2.23 | 0.077 |
| C | -7,244 | 134,221 | -0.05 | 0.959 |
| D | 4,536 | 134,221 | 0.03 | 0.974 |
| E | -6,842 | 134,221 | -0.05 | 0.961 |
| F | -514,976 | 134,221 | -3.84 | 0.012 |
| G | -532,049 | 134,221 | -3.96 | 0.011 |
| H | -225 | 134,221 | -0.00 | 0.999 |
| J | 300,377 | 134,221 | 2.24 | 0.075 |
| K | 542,896 | 134,221 | 4.04 | 0.010 |
| | S=536883 | $R^2$=92.5% | $R^2$(adj)=77.5% | |

To perform factor screening from the WIP model, a test of the significance of the individual regression coefficients was performed for each factor. It could be concluded that if the resulting p-value was less than 0.1 then the factor was significant. It should be noted that because a resolution III design was used, the main effects are confounded with two-way interactions. For example, if factor K was found to be statistically significant, it could possibly be due instead to the interaction of A & B. The decision to choose a model with such an aliasing structure is based on an economic tradeoff between experiment accuracy and resource availability.

For the purposes of this study factors having a p-value less than 0.1 were considered to be statistically significant; therefore, it can be concluded that factors having a p-value greater than 0.1 may be screened from the model. The test for significance of regression revealed that factors C, D, E, and H had a p-value greater than 0.1. These factors correspond to the increase in TSOP 256 loaders, the Increase in TSOP 54 Unloaders, The Increase in TSOP256 Unloaders, and the Increase in TSOP 256 Testers, respectively. Since these 4 factors do not have a significant effect on work in process, they were removed from consideration.

## 3 ROBUST DESIGN

The next section of this paper focuses on a discussion of the implementation of robust parameter design to the Infineon backend semiconductor manufacturing process. Based on the sensitivity analysis phase of the project, the six factors shown in Table 3 were believed to have a significant effect on both the WIP and cycle time.

Table 3: Significant Effects on WIP and Cycle Time

- Percent increase in burn in boards
- Increase in Ovens
- Percent hot lots
- Increase in TSOP54 Loaders
- Increase in TSOP54 testers
- Percent increase/decrease in product

The robust parameter design process is used to select levels of each controllable factor in the model to minimize the variance of the response while optimizing the response as much as possible. It is believed that the unexplained variance in the response is due to certain factors, which cannot be controlled in the process. This unexplained variance is inevitable; however, the variance is most likely sensitive to certain controllable factors in the model. This implies that the variance could at least be reduced by the proper selection of levels of the controllable factors (Ramberg et al. 1991).

Factors that effect the variance are called dispersion effects, while factors that affect the mean of the response are called location effects. In robust parameter design, it is believed that not all controllable factors are both dispersion

effects and location effects. Therefore, the variance can be minimized by properly adjusting the dispersion effects and the response can then be optimized by setting the factors, which are only location effects and not dispersion effects.

There are two main approaches to robust parameter design. A strict graphical approach and a dual response surface approach. The dual response surface response focuses on constructing regression models of the mean and the variance of the response. These models include both controllable and uncontrollable factors. From these regression models an estimate of the mean and variance of the response can be inferred at any point within the design space. These regression models make it easy to interpret the significant dispersion and location effects. Optimization can then be performed on the variance regression model and then on the mean response regression model with factors that are not significant dispersion effects. The dual response surface approach is also flexible in that it allows for constrained optimization of both regression models.

In the graphical analysis approach the mean and variance of the response are plotted along different points of a design space of controllable factors. Also, the mean and variance of the response can be plotted against each level of each controllable factor. The variance at each point in the design space of the controllable factors is measured by deviating the uncontrollable factors to specified levels. It should be noted that for experimental purposes that the uncontrollable factors can be adjusted to desired levels. Then from an analysis of each of these graphs the dispersion and location effects can be inferred. Optimization is performed by setting the dispersion effects to the levels that minimize variance and then setting the remaining location effects to levels that minimize response.

Both the dual response surface approach and the graphical approach to robust parameter design are effective methods. The dual response surface approach can only be employed when all factors of interest are continuous variables. At least four of the variables of interest to our study are qualitative in nature. Consequently, we used the graphical approach for much of our analysis.

### 3.1 Experimental Design

To perform a graphical analysis of the mean and variance a crossed array experimental design must be implemented. Many advocates of the Taguchi robust parameter design utilize the orthogonal array designs for both the inner and outer array of the crossed array design. A design is orthogonal if for any pair of columns, all combinations of factor levels occur and they occur an equal number of times. The method for robust design revolves around the use of orthogonal designs where an orthogonal array involving control variables is crossed with an orthogonal array for noise variables. The experimental design for the controllable factors lies in the inner array while the experimental design for the uncontrollable variables lies in the outer array. This design allows for an analysis of variance at the inner array points assuming that most of the variance is determined by the noise variables. The variances can then be compared at each of the experimental points and it then information can be inferred as to which factors have an effect on the variance (Montgomery 1990).

We will now proceed to describe the implementation of robust parameter design on the Infineon model. The first stage of robust parameter design consists of identifying uncontrollable factors that have an effect on the model variance. We chose the Cooling Time between the hot and cold test and the Time Between Unscheduled Failures for the MTX ovens as the two noise variables of consideration. It should be noted that there are other noise variables to consider, but the amount of experimental effort to analyze extra noise variables in the robust design process would be too time consuming.

The next step is to construct an appropriate crossed array for robust parameter design by selecting orthogonal array designs for both the controllable factors and for the uncontrollable factors. For this analysis a $L_8$ ($2^{6-3}$ factorial) orthogonal design was used for the inner array while a $2^2$ factorial design was used for the outer array points. This will provide an estimate for the mean response of all main effects and an estimate of variance for each setting of the main effects. This will result a 32 run experiment in which four runs, one for each setting of the noise variables, are performed at each of the eight design points of the controllable factors. The design is a resolution III design, meaning that all main effects are confounded with two-way interactions.

It should be noted that all of the degrees of freedom are dedicated to estimating the main effects. Due to the limited number of degrees of freedom available, the main effects are confounded with two-way interactions. This makes it impossible to determine conclusions as to which effect among an aliased group is contributing to variance in the response. Taguchi argues that we do not need to consider two-way factor interactions. He claims that it is possible to eliminate these interactions either by correctly specifying the response or design factors. However, we do not agree with this generalization and we believe that two-way interactions should be examined and they should be unaliased from main effects.

Folding over the $2^{6-3}$ into a $2^{6-2}$ x $2^2$ will result in a resolution IV design and it will then be possible to separate the confounding between the main effects and two-way interactions. Folding over is accomplished by adding an additional fraction to the design, which is a replicate of the original fraction and reversing all signs for the factors (Montgomery 1997). It should be noted however, that the two-way interactions remain aliased with other two-way interactions. So, one must be careful in analyzing the interaction plots for two way interactions and determine before hand which two-way factor interactions are aliased

with each other. Figure 3 is a graphical representation of the folded over $2^{6-2}$ design used for robust design.

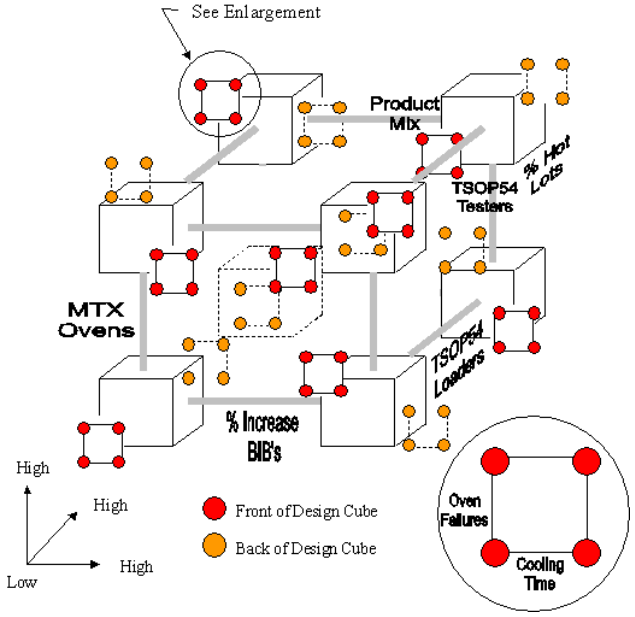


Figure 3: Folded-over $L_8$ Orthogonal Array

### 3.2 Analysis

The following section describes a number of graphical techniques useful for the analysis of Taguchi metamodels. During the course of this experiment two models were developed: one each for the mean and variance of WIP.

The Response-Scaled Run Plot is constructed by placing a scaled artifact at each point of the experimental design where an experimental trial will occur. Each artifact, typically a circle, represents the relative response of the model at that particular combination of design variables, i.e. in the case of circular artifacts, a circle having a larger diameter than that of another indicates that it, the larger circle, has a larger response (Barton 2000).

Applying these techniques to the Infineon model, we have developed a response-scaled run plot for the Work In Process given the design variables described previously. As a slight variation of the run-plots described above, we have chosen to plot both the response for mean and variance side-by-side.

Remember that the goal in robust design is to select independent variable settings that are "robust" or insensitive to random variations in the model; therefore, we would like to choose run conditions in which we maximize the signal-to-noise ratio (SNR). In other words, we would like to choose the point at which we minimize or maximize our response while minimizing the variance, or noise. As such, the Taguchi approach to parameter design really evolves into an "optimization" problem. By combining the responses for mean and variance into a singular plot, it is possible to determine the "optimal" run conditions that maximize the SNR (i.e. for the purposes of this model, the SNR is maximized when both the mean and variance are minimized – smaller the better).

Figure 4 is the Response-Scaled Run-Plot for the WIP response. The darker dots indicate the relative mean response and the lighter dots indicate the relative amount of variance observed for the run condition.

Analyzing the plot, it is easy to determine that the factors on the main or larger cube (%Increase of burn-in boards, Increase in TSOP54 Loaders, and Increase in MTX Ovens) have little if any effect on the WIP. This is indicated by the fact that there is insufficient change in the response on any of the eight cubes comprising the larger cube. The problem therefore reduces to an optimization of the region defined by the smaller cube.

Analyzing the smaller cube leads us to the conclusion that the TSOP54 Testers, %Hot Lots, and the Product Mix are the factors that most significantly affect the WIP. Observe that on all or at least most of the eight smaller cubes that as TSOP54 Testers is increased to its high level, both the mean and the variance of WIP decrease. On the other hand, as both the percentage of Hot Lots and the Product Mix are set to their high level, it can be observed that both the mean and the variance of WIP increase.

The fact that both the mean and the variance are minimized at the same combination of run settings is a very fortunate situation that occurs in our model. However this will not always be the case, and typically analysts will have to make decisions based on tradeoffs that exist between the mean response and the variability of the process.

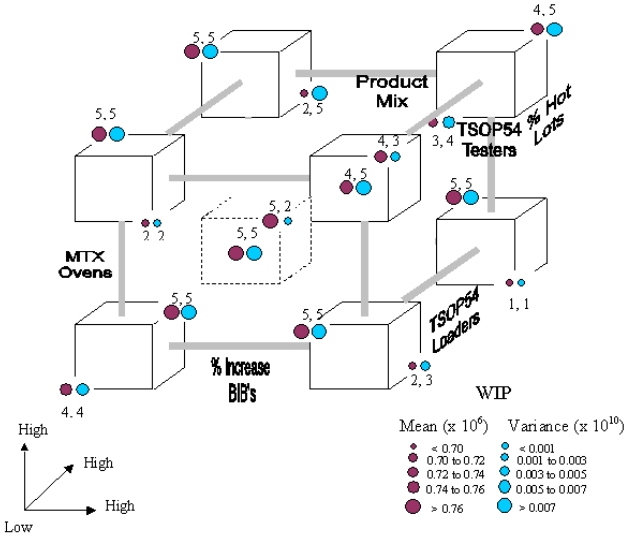


Figure 4: Response-Scaled Run Plot for WIP

The analysis of the Run-Plot for WIP suggests that the optimal configuration is at the high-level of TSOP54 Testers and the low-level for both the Product Mix and the %Hot Lots independent of the settings for %Increase burn-

in boards, TSOP54 Loaders, and MTX Ovens. However, it is important to note that it is not always possible run at the optimum conditions given market conditions, equipment availability, and premiums for special runs—this is where the true beauty of the Response-Scaled Run-Plot bears itself. Given a configuration it is possible to rapidly identify the settings at which the SNR can be maximized; thereby minimizing the overall impact to the system—assuming that the parameters are able to be reconfigured rapidly

### 3.3 Interactions

Based on an analysis of the interaction effects in the WIP model, we were able to identify possible significant interaction terms: AB, AC, BC, DE, DF, and EF. At least three of these terms are significant; however, because of the aliasing structure, it is impossible to tell which factors, within each aliasing group, are significant without further experimentation. Given that the aliasing structure of our design is I = ABEF = BCDE = ACDF, it is possible to determine the aliasing structure of all two-factor by two-factor designs (Montgomery 1997). Some of these significant interaction terms are aliased with each other in our design making it impossible to tell which factor, within each aliasing group, is significant without further explanation.

### 3.4 Conclusions

Through graphical analysis, we have concluded that the '%Increase of burn-in boards', 'Increase in TSOP54 Loaders', and 'Increase in MTX Ovens' have little if any effect on WIP or cycle time. The problem, therefore, reduces to an optimization of the region defined by the TSOP54 Testers, %Hot Lots, and the Product Mix.

Based on the graphical analysis it is evident that as TSOP54 Testers is increased to its high level, both the mean and the variance of WIP and cycle time decrease. On the other hand, as both the percentage of Hot Lots and the Product Mix are set to their high level, it can be observed that both the mean and the variance of WIP (and cycle time) increase.

The analysis suggests that the optimal configuration is at the high-level of TSOP54 Testers and the low-level for both the Product Mix and the %Hot Lots independent of the settings for %Increase Burn-In Boards, TSOP54 Loaders, and MTX Ovens. It is not always possible run at the optimum conditions due to market conditions, equipment availability, and premiums for special runs. However, this is where the use of the Response-Scaled Run Plot has its advantages. From any configuration it is possible to rapidly identify the settings at which the SNR can be maximized; thereby, minimizing the overall impact to the system.

## 4 SUMMARY

We have used graphical methods for conduction a robust design study on a semiconductor backend Manufacturing process. After a factor screening step, we were able to reduce the number of factors to less than ten, making the graphical approach practical. The methodology is easy to apply and to interpret, even in the presence of qualitative variables.

## ACKNOWLEDGMENTS

We would like to acknowledge Steven Brown, Juergen Potoradi, and employees of Infineon for their insight and assistance. We would also like to thank Frank Chance for his assistance with Factory Explorer.

## REFERENCES

Barton, Russell R. 1999. Graphical Methods for the Design of Experiments. New York: Springer.

Barton, Russell R. 2000. Using Simulation Models for Experimental Design. Unpublished Lecture Notes, Harold & Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA.

Montgomery, Douglas C. 1997. Design and Analysis of Experiments. New York: John Wiley and Sons.

Montgomery, Douglas C. 1990. Using Fractional Factorial Designs for Robust Process Development. Quality Engineering 3 (1): 193-204.

Ramberg, John S., et al. 1991. Designing Simulation Experiments: Taguchi Methods and Response Surface Metamodels. In Proceedings of the 1991 Winter Simulation Conference, ed. B. L. Nelson, W. D. Kelton, G. M. Clark, 167-176. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Devore, Jay L.. 1995. Probability and Statistics for Engineering and the Sciences, 4th ed. Belmont: Duxbury Press.

## AUTHOR BIOGRAPHIES

**SCOTT L. ROSEN** is a Ph.D. candidate in Industrial Engineering and Operations Research at the Pennsylvania State University. He received his B.S. Industrial Engineering from Lehigh University in 1998 and his M.S. from Pennsylvania State University in 2000. His current interests include simulation-based optimization and scheduling. He is currently a member of INFORMS. His email address is <slr209@psu.edu>.

**CHAD A. GEIST** is a graduate student at The Pennsylvania State University pursuing an M.S. in Industrial and Manufacturing Engineering. He received his B.S. in Industrial and Manufacturing Engineering from The Penn-

dustrial and Manufacturing Engineering from The Pennsylvania State University in 1999. His current interests include process planning, simulation of manufacturing systems, computer integrated manufacturing, CAD/CAM and quality control. He is currently a member of SME and SAE. His email address is <cag141@psu.edu>.

**DANIEL A. FINKE** is a graduate student at The Pennsylvania State University pursuing an M.S. in Industrial Engineering and Operations Research. He received his B.S. in Industrial Engineering from New Mexico State University in 2000. His current interests include simulation-based optimization and decision improvement. His email address is <daf903@psu.edu>.

**JYOTIRMAYA NANDA** is a graduate student at The Pennsylvania State University pursuing an M.S. in Industrial Engineering. He received his B.S. in Mechanical Engineering from at Visvesvaraya Regional College of Engineering in 1998. His current interests include internet-base mass customization and simulation-based design. His email address is <jnanda@psu.edu>.

**RUSSELL R. BARTON** is a Professor in the Harold and Inge Marcus Department of Industrial and Manufacturing Engineering at Penn State. He received a B.S. in Electrical Engineering from Princeton University and M.S. and Ph.D. degrees in Operations Research from Cornell University. At Penn State, he has worked to increase the practice component of engineering education. He is Secretary-Treasurer for the Informs College on Simulation and he was Proceedings Co-editor for the 2000 Winter Simulation Conference. His current interests include graphical methods for experiment design, design of experiments and metamodeling methods applied to simulation models, and statistical models of product and process behavior. His hobbies include travel, golf, and a 1937 Pontiac. His email address is <rbarton@psu.edu> and his web page address is <http://www.ie.psu.edu/people/faculty/barton.htm>.