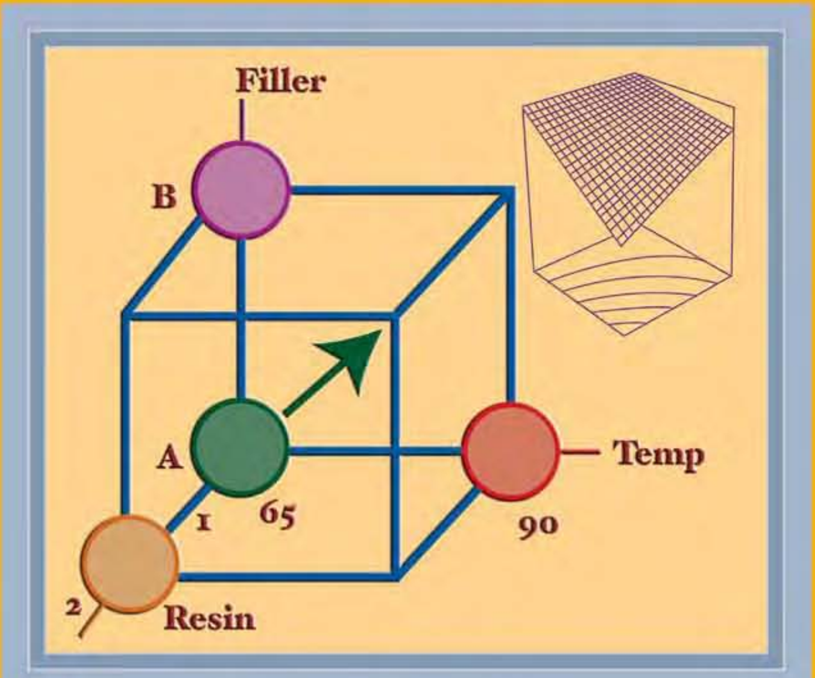
# Pharmaceutical and Medical Device Validation by Experimental Design



Edited by

Lynn D. Torbeck

informa
healthcare

# Pharmaceutical and Medical Device Validation by Experimental Design

# Pharmaceutical and Medical Device Validation by Experimental Design

Edited by

**Lynn D. Torbeck**

*Torbeck & Associates, Inc.*
*Evanston, Illinois, USA*

informa
healthcare

New York London

Informa Healthcare USA, Inc.
52 Vanderbilt Avenue
New York, NY 10017

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 1-4200-5569-0 (Hardcover)
International Standard Book Number-13: 978-1-4200-5569-6 (Hardcover)

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequence of their use.

**Library of Congress Cataloging-in-Publication Data**

Pharmaceutical and medical device validation by experimental design /
edited by Lynn D. Torbeck.
p. ; cm.
Includes bibliographical references and index.
ISBN-13: 978-1-4200-5569-6 (hb : alk. paper)
ISBN-10: 1-4200-5569-0 (hb : alk. paper) 1. Pharmaceutical technology–Validity–Research–Methodology. 2. Experimental design.
I. Torbeck, Lynn D.
[DNLM: 1. Technology, Pharmaceutical–methods. 2. Multivariate Analysis. 3. Quality Control. 4. Reproducibility of Results. 5. Research Design. 6. Technology, Pharmaceutical–standards. QV 778 P5353995 2007]

RS192.P468 2007
615'.19–dc22 2007001418

**Visit the Informa Web site at**
**www.informa.com**

**and the Informa Healthcare Web site at**
**www.informahealthcare.com**

# Foreword

Imagine the following scenario: you have a seemingly impossible quality problem to solve. A new lyophilized product your company is about to launch is apparently unstable at commercial scale. Management is breathing down your neck for an immediate solution, since the research and development scale-up data indicated the formulation is stable. A crisis team was formed and has been working on a solution for almost six months.

Seeing the difficulty, a colleague suggests a different approach to looking at the problem, calling together a subteam of four people that are the most knowledgeable about the production process. They use a few simple tools and determine the most probable cause of the problem in a couple of hours using existing data. The resulting action plan corrects the problem, stable lyophilized lots are produced, and the product is launched on schedule. Does this sound too good to be true?

Fortunately, for those facing similar seemingly insuperable problems, the story is actually true. Not only is this specific story true, but also there are many more like it. Virtually every major biopharmaceutical company in the world has its own case study. These solutions were achieved thanks to a few simple process analysis tools and a systematic structured approach to visualizing data-rich experiments.

How do I know they work? Because I have personally used them, I have taught people to use them, and I have seen the results myself first hand.

Read this book. Buy copies for the people in your company who work on difficult problems and train them in the methodology. Use the techniques to help find your own solutions to nearly impossible problems. Do not miss a chance to use the most powerful toolkit I have seen used in my 35 years in the industry. If you do not use these tools, you are also missing the opportunity to bring robust life saving therapeutics, as quickly as possible, to the people we really work for—the patients and their families.

## THE APPROACH

DOE is an acronym for design of experiments, also called experimental design or multifactor experiments. These experimental approaches provide rich veins of data that can be mined for information that cannot be found any other way.

The reason for design in the description is that DOE experiments must be scientifically sound and statistically valid, especially for use in the highly regulated biopharmaceutical industry. Scientific soundness is the cornerstone of the scientific peer review process, and (read world regulatory authority) Food and Drug Administration review and approval of such studies in support of product license approval. Statistical validity is necessary to ensure the integrity of the experiments and to appropriately interpret the significance of the data. In real terms, these are the same science and statistics that are applied to clinical trials; now the application is being extended to production process monitoring, quality control testing, validation and process analytical technology or PAT. The DOE concepts are also valuable tools capable of being used to exploit existing data and help solve seemingly impossible product quality problems.

Scientists and managers are traditionally taught that experiments should only change one factor at a time (e.g., temperature), holding other factors constant (time, pH, etc.), in order to explain the individual effect on product quality. By contrast, the design of experiments trials are carefully worked out to assess many possible factors (time, temperature, and pH) at various levels (one hour vs. two hours at 25°C vs. 30°C at pH 6.5 vs. pH 7.5) to determine which one, or combination, has the greatest effect on the product's quality characteristics, such as yield, impurities, and viral reduction. DOE measures not only the single factor effects, but also the cumulative and interaction effects of all the factors investigated on product quality. Most important is the fact that DOE is the only way to see interaction effects; one factor at a time experiments give you just that, one effect for each factor.

## INTERACTION EFFECTS

DOE is the only technique that enables scientists and managers to find, see, and use interaction effects that can improve product quality and yield or help set process boundaries to prevent failure. The well entrenched views that only one factor at a time can be studied and the widely held management maxim that "if it ain't broke, don't fix it" are not only wrong but, in some cases, dangerous; a process can drift into failure, or periodically have an "unexplainable" failure, due to interaction effects. DOE permits scientists to conduct small scale, low cost process improvement experiments to model large scale, "unbroken" commercial processes without endangering current production runs and product supply. These same experiments can also generate data that drives continuous process improvement (read "make more profitable") of a process that "ain't broke."

## LEAST COST/FASTEST ROUTE

DOE is the least costly way to collect data. First, the basis for performing a series of designed experiments is usually reviewing the existing process database; the longer a process has run the richer the possibilities. Second, the experiments are usually conducted using a small scale process simulation—a much less expensive approach than tying up the production line. Third, the scale and cost of down-sized runs permits a relatively larger number of runs in a relatively short period of time. The resulting body of data provides a roadmap for process improvements that can be verified by additional small scale or scaled up experiments prior to full scale transition that meets ongoing production requirements.

## VALIDATION

DOE's data will define critical process parameters for validation. The results of designed experiments are the identification of the individual parameters and those parameter interactions that have the most effect on product quality and yield. These are the critical process parameters that need to be assessed during process validation. Just as important, DOE identifies those parameters that do not impact product quality. These do not need to be validated but rather monitored and controlled within their defined process ranges. The savings of validating the DOE defined critical process parameters versus validating "everything we think may be critical" is substantial in time, monetary, and human resource terms. Furthermore, given the scientific soundness and statistical validity of DOE results, processes validated using properly applied DOE principles are nearly bullet proof with the world's regulatory agencies.

## SUMMARY

DOE is the fastest route to a profitable, reliable, robust, validated process. DOE's requirement of a rigorous design methodology that passes peer review with the scientists most knowledgeable about the process ensures scientific soundness. The depth of DOE's statistical foundation that enables the measurement of multiple effects and interactions in a single set of experiments proves DOE's statistical validity. DOE is also a resource conservator, since it requires less time and provides data quicker than single factor at a time experiments. The proven statistical validity of the DOE technique guarantees accurate analysis of the data and valuable information for management to turn data analysis and interpretation into action.

*Ronald C. Branning*
*Genentech, Inc.*
*South San Francisco,*
*California, U.S.A.*

# Preface

Designed experiments were shown to be useful for validation almost 30 years ago. The famous chemist/statistician, W. J. Youden, illustrated the use of Plackett–Burman designs for ruggedness testing in the 1975 book, *Statistical Manual of the Association of Official Analytical Chemists* (1). In that short introduction, he noted that "… if the program is carefully laid out, a surprisingly small amount of work suffices." The example given showed seven factors varied in eight runs. Each of the seven factors is estimated using all eight data values. This is the equivalent of having 56 data points, but only spending the money to buy eight. This efficiency is very attractive to laboratory managers who are charged with validating many analytical methods in a short time. Note that ruggedness testing is still required today by the Food and Drug Administration in the 1987 *Guideline for Submitting Samples and Analytical Data for Method Validation* (2).

This editor gave his first talk on using designed experiments in 1978 in response to the then draft Current Good Manufacturing Practices. In 1984, Chao, St. Forbes, Johnson, and von Doehren (3) gave an example of granulation validation using a $2^3$ full factorial. Since then, there have been many journal articles and conference talks showing the application of designed experiments to validation. A recent addition is the text by Lewis et al., *Pharmaceutical Experimental Design* (4).

Yet, for all of the dispersed information and examples, there still is unease in the pharmaceutical industry about using designed experiments in general or for validation specifically. Typical questions include, "What will the Food and Drug Administration say if I use a designed experiment for validation?" or "Won't that take a lot longer and cost a lot more money?" In answer to the first question, the Food and Drug Administration requires the industry to use carefully designed experiments in clinical and animal studies. The well-regarded double blind clinical trial is a designed experiment. These are usually done exactly by the book. It is hard to see the Food and Drug Administration objecting to the use of designed experiments in validation when it is required in other areas. Further, in 1985, Ed Fry, while with the Food and Drug Administration, said in his excellent article "Then, the processes that cause variability … must be identified. Experiments are conducted (that is validation runs) to ensure that factors that would cause variability, are under control …" (5).

This book is the answer to the second question. Designed experiments are the most scientific, the most efficient, and the most cost effective way we know how to collect data. Designed experiments need not be complicated or statistically complex. Realistic case studies and examples illustrate the use of design of experiments for validation. The use of graphics illustrate the designs and results. Formulas are minimized and used only where necessary. Where possible, a step-by-step approach or procedure is given. Detailed protocols and reports with realistic data are given where appropriate. This book succeeds if the reader feels that it is obvious that design of experiments is the most logical and rational approach to use.

A variety of examples and case studies are given to show the wide range of application. Assay and bioassay validation, process and equipment validation, etc. Not all cases are "end point" validation, but are further up-stream, and part of the life cycle validation discussed by Chapman (6).

Each chapter stands alone. It is not necessary to read them in sequence. The reader is encouraged to delve into the chapters that seem most applicable and interesting. The text is intended to be a learn-by-doing-by-example. Find an example close to a intended project and mimic the approach.

It is hoped by the authors and the editor that this text will encourage people to use designed experiments for validation and learn first hand the benefits.

*Lynn D. Torbeck*

## REFERENCES

1. Youden WJ. Statistical Manual of the Association of Official Analytical Chemists. Arlington, VA: AOAC, 1975.
2. FDA, Guideline for Submitting Samples and Analytical Data for Method Validation, Rockville, Maryland, 1987.
3. Chao A, Forbes S, Johnson, R, von Doehren P. Prospective process validation. In: Loftus B, Nash R, eds. Pharmaceutical Process Validation. New York: Marcel Dekker, 1984:125–148.
4. Lewis G, Mathieu D, Phan-Tan-Luu R. Pharmaceutical Experimental Design. New York: Marcel Dekker, 1999.
5. Fry E. The FDA's viewpoint. Drug and Cosmetic Industry 1985; 137(1):46–51.
6. Chapman K. A history of validation in the United States: part 1. Pharmaceutical Technology October 1991:82–96.

# Contents

# Contributors

**Ronald C. Branning** Genentech, Inc., South San Francisco, California, U.S.A.

**Robert F. Dillard** Biostatistics and Data Management, Takeda Pharmaceuticals North America, Inc., Lincolnshire, Illinois, U.S.A.

**T. Lynn Eudey** Department of Statistics, California State University East Bay, Hayward, California, U.S.A.

**Jeffrey T. Field** J. T. Field Consulting Services, LLC, Woodbury, Connecticut, U.S.A.

**Sourav Kundu** Technical Operations, Aventis Behring, Bradley, Illinois, U.S.A.

**David M. Lansky** Lansky Consulting, LLC, d/b/a Precision Bioassay, Burlington, Vermont, U.S.A.

**Thomas D. Murphy** T. D. Murphy Statistical Consulting, LLC, Morristown, New Jersey, U.S.A.

**Daniel R. Pilipauskas** Global Manufacturing Services, Pfizer, Inc., New York, New York, U.S.A.

**Wayne A. Taylor** Taylor Enterprises, Inc., Libertyville, Illinois, U.S.A.

**Lynn D. Torbeck** Torbeck & Associates, Inc., Evanston, Illinois, U.S.A.

1

# Designing Experiments for Validation of Quantitative Methods

**T. LYNN EUDEY**

Department of Statistics
California State University East Bay
Hayward, California, U.S.A.

## INTRODUCTION AND SCOPE

Quantitative methods are assays that result in meaningful numeric measurements for a characteristic of a product. Quantitative methods are used in assessing whether final product meets specifications. They are also used to measure product quality (or quantity) in various stages of manufacturing and the results are often used in quality control charts. Validation is an objective process used to determine whether a quantitative method is performing as expected and is appropriate for its intended use. This chapter provides the motivation behind validation, some terms and definitions used in validation, a consolidated statistically sound approach to validation, along with appropriate statistical analysis, and reporting of validation results. A hypothetical but realistic example is presented and is used to illustrate the validation process.

Motivation and some of the logistics of validation are presented here in the introductory section of the chapter. An example protocol is presented in the second section, followed by a section of terms and definitions. Design of experiments, presented in the fourth section, is used to ensure that the validation experiments represent the populations of all "runs" of the method that are being validated. Pragmatic limitations are discussed in this section.

The fifth section is a continuation of the example and contains a hypothetical data set with an analysis. The sixth section discusses statistical analyses and illustrates an analysis of the example validation data.

In this chapter, "method" refers to the procedure of interest—the method being validated. The term "assay" is defined as a single execution of this method, possibly abstract, while "run" refers to an actual single execution of the method. Often a procedure will call for multiple measures within an assay; these are referred to as "replicates." The reportable value of an assay could be the result of one replicate or the average of multiple replicates (here, this is in the abstract sense; it is the formula, or formulae, which will be used to calculate the reported value). If the reportable value is defined as the average of three replicates, then the reported value would be the average of the observed three replicate values from one run of the method.

## Why Validate?

Validation can be a method of quantifying the performance of a process; in this case measuring the performance of a quantitative method. In 1985, E. M. Fry wrote: "Validation has a quantitative aspect—it's not just that you demonstrate that a process does what it purports to do; you actually have to measure how well its does that … then, the processes that cause variability … must be identified. Experiments are conducted (that is, validation runs) to ensure that factors that would cause variability, are under control (1)."

The process in question here is a quantitative method. The question becomes: "How well does the method measure the parameter that it purports to be measuring?" Validation provides an objective measure of a method's performance. Using samples with a known (or at least previously measured) value of a product parameter, validation can provide useful information about accuracy, precision, linearity, and other characteristics of the method's performance outside of its daily use on unknown samples. In addition, validation can be used to identify sources of undesired variability.

Validation of a method involves running assays on aliquots from the same sample a number of times and ideally is done over a well-balanced design of all the external factors effecting performance. If, for example, more than one machine could be used for the method in question then, ideally, all machines should be tested in the validation. In a similar manner, the validation should be run over a number of days (or otherwise changing environmental conditions) to show the method is rugged to typical environmental changes. Additionally, more than one analyst should run the validation assays if that is the normal practice or if the analyst position has a high turnover.

## What Are We Measuring?

Validation is the process of measuring the performance of a previously developed quantitative method. The characteristics measured to assess performance are defined and explained in detail later in this chapter. Typically, for a quantitative method the characteristics of specificity, accuracy, precision, and linearity are measured. The range of the method can be defined, or verified, by the region where the aforementioned characteristics are acceptable. Although their determination is more appropriately a development task, the characteristics of robustness, limit of detection, and limit of quantitation can also be verified during validation.

The discussion here centers on the quantitative characteristics of accuracy, precision, and linearity. Design of experiments is also useful in supporting the validation of specificity and studying robustness (refer to section on Validation Terms and Definitions).

## Development Versus Validation

Although the validation approach can be used in development, for simplification the author assumes that the method is fully developed before validation. Validation is an objective way of verifying that the development stage has been successful and the method is performing to expectations. Typically, the range of the method as well as the method's input parameters will be defined during development and verified during validation.

Robustness of a method is defined as "a measure of its capacity to remain unaffected by small but deliberate variations in method parameters and provides an indication of its reliability during normal usage" [International Conference Harmonisation (ICH) Guideline Q2A (2)]. Intervals for method parameters (e.g., ranges for elements or conditions of sample preparation) should be honed in the development stage to ensure consistency in assay results and may be verified during validation. Ranges for factors internal to the method such as incubation time and temperature are also determined during development. The standard operating procedure (SOP) should state ranges for method factors where assay performance is fairly consistent or robust. Ideally, the range is specified as target (low and high) so that the operator is aware of the target for the parameter, as well as how much the parameter can vary, and the method will still have similar assay performance. For example, a method with an incubation time specified 25 minutes (24.5, 25.5) is less robust with respect to incubation time than a method with an incubation time specified 25 minutes (20, 30).

Experimental design can play a key role in finding these ranges of robustness during method development, but validation is a step taken after the appropriate ranges are set. Thus, for validation, each assay is run per the SOP and the internal factors are generally considered constants. If there is an acceptable range for a specific input parameter then validation can be used to verify that the method is still robust over the SOP specified range.

## What Is Needed? Guidelines for Validation

Agencies, such as the Food and Drug Administration (FDA), United States Pharmacopoeia (USP), and ICH issue guidelines for method validations. Torbeck (3) has provided a comprehensive list of guidelines and references in his course on assay validation. For purposes of this chapter the ICH guidelines Q2A and Q2B (2,4) will be used. There are some minor differences in nomenclature between the different guidelines, and some of these will be explained in the short section on terms and definitions.

The purpose of a method validation is to show that the method performs to the expectations of the user. A priori expectations need to be set out in the form of acceptance criteria. These must be formally stated using a preapproved protocol. In addition to setting down a priori acceptance criteria for the method's performance characteristics, the protocol is a document written to specify how the validation will be run, the layout of the experimental design, the form of documentation for training and for execution of the validation experiments, how the data will be collected and analyzed, and additionally provides a structure for writing a final report. A hypothetical protocol and highlights of a statistical report will be used to demonstrate this process. It is important to remember that the intent of validation is to show that a method is acceptable for its intended use. Keep in mind that *validation is not an exploratory or development process.*

## VALIDATION PROTOCOL

### Protocol Content

A validation protocol states the purpose of the validation method, the intended substance being tested, the definition of the reportable value, the validation approach, the specific directions for conducting the validation assays, the statistical approach for analyzing the resulting data, and the nonambiguous acceptance criteria. The protocol should allow no room for ambiguity in the execution of the protocol or in the acceptance criteria. The content of a protocol is described next. There are many ways to format a protocol; herein are suggestions for the content of a protocol.

The following is a hypothetical example of a validation protocol. For purposes of this example we will assume that SOP 123 is the procedure used to measure mass (in $\mu$g) using a calibration curve from a standard solution (all samples are pipetted to 1 $\mu$L; hence, we are actually measuring concentration, but volume is considered constant at 1 $\mu$L, so the term "mass" will be used for the measurement). Validation performance characteristics are briefly defined in the context of a protocol. These terms are discussed in more detail in the following section on terms and definitions. Square brackets [ ] delineate general content of a section or parenthetic remarks.

The quantitative method characteristics to be validated will depend on the nature of the method itself. The reader should refer to the table in ICH guideline Q2A, which "lists those validation characteristics regarded as the most important for the validation of different types of analytical procedures." The design shown in the example validation matrix allows simultaneous assessment of accuracy, precision, and linearity.

Ideally, the expected masses are obtained by an alternative physical methodology that is highly accurate and precise. In the case of using a commercial standard, the vendor should provide statistics on the accuracy and precision of the method used to obtain/measure the stated label concentration. When using a previously released lot, the expected masses will be based on the certificate of analysis concentration for the lot of *Analyte B* used in this validation. In the latter case, the measurement of accuracy is relative to the historical certificate of analysis value. The validation protocol needs to state that the accuracy measurement is relative to the historical value rather than to an independently obtained measurement.

## EXAMPLE VALIDATION PROTOCOL

This section presents an example validation protocol. As mentioned before, square brackets [ ] are used to discuss content in general.

### PROTOCOL FOR THE VALIDATION OF SOP 123 FOR USE IN THE DETERMINATION OF MASS OF PRODUCT W

[The title page, or some other page prefacing the protocol, must include signatures, with dates of signing, of all appropriate personnel to show approval of the protocol, including, but not limited to, the author of the protocol, the personnel responsible for the execution of the validation assays and collection of the data, the personnel responsible for the data analysis, as well as appropriate quality assurance and management.]

### §1. Purpose and Scope

The purpose of this protocol is to define the method performance characteristics to be observed, the design and execution of the validation experiments, the data analysis, and the acceptance criteria for the validation of SOP 123 for use on Product W.

[The scope of the document defines the specific product and method combination being validated. It also states the department responsible for running the assays and the facility (or facilities) where the method is being validated. If more than one facility is involved then reproducibility of the method between laboratories needs to be validated.]

### §2. Background

[This section describes the background of the use of SOP 123 and a brief description of the method.] The current range of this method for Product W is 50 $\mu$g to 150 $\mu$g.

### §3. Validation Approach

[This section describes the general approach in validating the method, lists and defines the validation characteristics to be evaluated (refer to Section III. Terms and Definitions).]

To validate SOP 123 for measuring mass of Product W, the quantitative method performance characteristics of accuracy, precision, linearity, and range will be assessed using the validation assays shown in the design matrix over two days and using two operators. As per ICH guideline Q2A, the validation experiments will be run from 40 $\mu$g to 180 $\mu$g. The test lot will be diluted out and concentrated up to specific expected masses using the mass cited on certificate of analysis for the lot of Product W selected for the validation. The points on the Product W curve will be as follows: 40 $\mu$g, 50 $\mu$g, 70 $\mu$g, 90 $\mu$g, 110 $\mu$g, 130 $\mu$g, 150 $\mu$g, and 180 $\mu$g.

Accuracy in percent recovery is defined by the ratio of the observed test lot mass to the certificate of analysis mass for the test lot. Thus, accuracy is relative to the previously measured mass and not by comparison to an objective measure of mass of the test lot.

Precision of the method is defined by the sum of intermediate precision (interday, interoperator, and interassay precision) and repeatability.

Linearity of the method is assessed by the linear regression statistics of observed mass against expected mass.

Current range of the method will be verified by acceptable Accuracy, Precision, and Linearity (see Acceptance Criteria).

[Other procedural details such as test lot preparation to aliquot all the samples needed for the validation experiments should also be described briefly in this section of the protocol.] Referring to SOP 123, the reportable value is defined as the average of the three replicates. For this validation, accuracy, precision, and linearity of the replicate values will be assessed; consequently, the precision of the results reported is at the replicate level not the reportable value level. [Note this practice is done for two reasons: first, to conserve resources, and second, when it is not possible to repeat an additional assay with exactly the same sample preparation. The repeatability component of the precision is defined as within-assay variance. Using the replicate values yields a within-assay variance.]

### §4. Responsibilities and Documentation

[This section, or sections, of the protocol describes the responsibilities of the pharmaceutical or biotechnology company and the departments within the company in carrying out a validation as well as listing the documentation needed for reference including SOP 123.]

### §5. Training Verification and Documentation

[Record on a Training Record that the personnel executing the validation experiments are adequately trained in compliance with company quality policies. Additionally, document that these personnel have been trained on the validation protocol.]

### §6. Test Work

[This section describes the execution of the validation experiments, how this execution may differ from the SOP, how and where the data will be recorded, as well as specifying how each of the validation characteristics will be evaluated.] In assays A, B, C, and D use the method of SOP 123 to determine triplicate measures of mass for Product W. For each point on the Product W curve (40 $\mu$g, 50 $\mu$g, 70 $\mu$g, 90 $\mu$g, 110 $\mu$g, 130 $\mu$g, 150 $\mu$g, and 180 $\mu$g) record each triplicate measure on the test report form.

## §6.1. Accuracy

### *§6.1.1. Data Analysis*

For each point on the Product W curve, divide the observed mass by the expected mass and express as a percent. For each expected mass, the average percent accuracy will be reported with a 95% confidence interval [as per the ICH guideline Q2B].

### *§6.1.2. Acceptance Criteria*

[Acceptance criteria appropriate to the use of the method for accuracy are stated in the protocol before the data analysis. Often the acceptance criteria are based on design criteria for the development of the method or development data.]

For each mass in the tested range observed, average accuracy must be between 90% and 110%. [These are hypothetical acceptance criteria.]

## §6.2. Precision

### *§6.2.1. Data Analysis*

For each mass on the Product W curve, the data will be analyzed by variance components analysis (VCA) to estimate the precision components due to InterOperator, InterDay, InterAssay, and Repeatability. For each expected mass, the precision components will be recorded as a variance, standard deviation, and percent of total precision. Nonzero components will be reported with a two-sided 95% confidence interval for the standard deviation. Within each mass, total variance is defined as the sum of the variance components. Total precision will be expressed as a variance, standard deviation, and coefficient of variation (%CV) also called the percent relative standard deviation or %RSD.

### *§6.2.2. Acceptance Criteria*

[Acceptance criteria appropriate to the use of the method for precision are stated in the protocol before the data analysis.

Often, the acceptance criteria are based on design criteria for the development of the method or development data.]

For each mass in the tested range observed, the total %CV must be less than 10%. [These are hypothetical acceptance criteria. Note that the precision acceptance criterion can be on the standard deviation or some other metric of variability, preferably a metric that is somewhat consistent in value over the range of the assay.]

## §6.3. Linearity

### *§6.3.1. Data Analysis*

Within each of the assays A, B, C, and D, least squares linear regression of observed mass will be regressed on expected mass. The linear regression statistics of intercept, slope, correlation coefficient ($r$), coefficient of determination ($r^2$), sum of squares error, and root mean square error will be reported. Lack-of-fit analysis will be performed and reported. For each assay, scatter plots of the data and the least squares regression line will be presented.

### *§6.3.2. Acceptance Criteria*

For each assay, the coefficient of determination must be greater than 0.975. [These are hypothetical acceptance criteria. Other metrics for linearity could be used for the acceptance criterion. Note that for precise assays, a significant lack-of-fit may not be a meaningful lack-of-fit, due to a slight but consistent curve or other artifact in the data. If lack-of-fit is used for the acceptance criterion, the requirement should be placed on absolute deviation or percent departure from linearity rather than statistical significance. Often the acceptance criteria are based on design criteria for the development of the method or development data.]

## §6.4. Range

Range is defined by the interval where the method has demonstrated acceptable accuracy, precision, and linearity.

## §6.5. Other Performance Characteristics

[Other method performance characteristics should be addressed as appropriate. The reader is encouraged to refer to ICH Q2A to determine which performance characteristics need to be evaluated in the validation. Depending on the characteristics addressed, the validation may need more than one design of experiments.]

## §7. Validation Matrix

[A matrix of assays is presented to display the study design over the appropriate factors. For this example (Table 1), the factors are Operator, Day, and Assay. For this example, Assay is defined as the interaction between Operator and Day. Note that this matrix shows a hierarchical, or nested, model as Assay (sample preparation) is uniquely defined by Operator and Day.]

## §8. Attachments and Forms

[This section contains forms for documenting that all training requirements have been met, sample preparation instructions as necessary to the execution of the validation experiments, and data recording forms. Places for signatures for performance and review should be included on each form as appropriate to company quality and documentation standards. For examples, see Tables 2 and 3.]

**Table 1** Matrix of Validation Assays

| Assay | Day | Operator | Replicates |
|---|---|---|---|
| A | 1 | I | 3 |
| B | 1 | II | 3 |
| C | 2 | I | 3 |
| D | 2 | II | 3 |

Four assays with three replicates each. Day 1 is a separate day than Day 2.

**Table 2** Training Verification to Include Training on Standard Operating Procedure (SOP) 123 and Training on This Validation

| Name | Department | List SOP | Training record complete | Verified by—initials/date |
|---|---|---|---|---|
| | | | | |

Add a section for comments, as well as a place for signatures for the personnel who reviewed the training records and conducted the validation training in accordance with company quality standards for training documentation.

**Table 3** Assay Recording Sheet, One for Each Assay

Assay Data Recording Sheet for Validation of SOP 123
Copy 1 for Each Assay

Date of assay: ____________________
Assay validation ID: ________________
Operator: _______________________
Sample lot number: ________________

| | | Response per replicate (absorbance, light units, etc.) | | | Replicate calibrated value of mass (μg) | | |
|---|---|---|---|---|---|---|---|
| Sample | Expected mass (μg) | 1 | 2 | 3 | 1 | 2 | 3 |
| Standard | 40 | | | | | | |
| | 50 | | | | | | |
| | 70 | | | | | | |
| | 90 | | | | | | |
| | 110 | | | | | | |
| | 130 | | | | | | |
| | 150 | | | | | | |
| | 180 | | | | | | |
| Test lot | 40 | | | | | | |
| | 50 | | | | | | |
| | 70 | | | | | | |
| | 90 | | | | | | |
| | 110 | | | | | | |
| | 130 | | | | | | |
| | 150 | | | | | | |
| | 180 | | | | | | |

Include rows/columns for standard curve and other data normally recorded during the execution of standard operating procedure (SOP) 123.

## TERMS AND DEFINITIONS

The following gives definitions of method validation characteristics following the references (2,4) ICH Q2A "Text on Validation of Analytical Procedures" and ICH Q2B "Validation of Analytical Procedures: Methodology." ICH Q2A identifies the validation characteristics that should be evaluated for a variety of analytical methods. ICH Q2B presents guidelines for carrying out the validation of these characteristics.

The validation characteristics of accuracy, precision, and linearity were presented in the example protocol of the previous section.

## Accuracy

Accuracy is a measure of how close to truth a method is in its measurement of a product parameter. In statistical terms, accuracy measures the bias of the method relative to a standard. As accuracy is a relative measurement, we need a definition of "true" or expected value. Often, there is no "gold standard" or independent measurement of the product parameter. Then, it may be appropriate to use a historical measurement of the same sample or a within-method control for comparison. This must be accounted for in the design of experiments to be conducted for the validation and spelled out in the protocol. Accuracy is measured by the observed value of the method relative to an expected value for that observation. Accuracy in percent can be calculated as ratio of observed to expected results or as a bias of the ratio of the difference between observed and expected to the expected result. For example, suppose that a standard one-pound brick of gold is measured on a scale 10 times and the average of these 10 weights is 9.99 lbs. Then calculating accuracy as a ratio, the accuracy of the scale can be estimated at $(9.99/10) \times 100\% = 99.90\%$. Calculating the accuracy as a bias then $[(9.99 - 10)/10] \times 100\% = -0.10\%$ is the estimated bias. In the first approach ideal accuracy is 100%, and in the second calculation ideal bias is 0%.

## Precision and Ruggedness

Precision is a measure of how variable a method is in its measurement of a product parameter under normal usage. In statistical terms, precision is measured by the variance of the method. Additionally, the typical sources of variability are accounted for in assessing precision; these are random factors external to the method, such as analysts, days, and changing assay "hardware" (e.g., gels or columns). Variance components measure the contribution of each of these external factors to the variability of the method's results on the same sample of product. To estimate the variance components, the design of validation experiments must be balanced with respect to these factors.

Precision components are defined at three levels: reproducibility, intermediate precision, and repeatability. Reproducibility is the variability of the method between laboratories or facilities. However, as a laboratory is not "randomly selected" from a large population of facilities, laboratory is a fixed effect. Consequently, the assessment of reproducibility is a question of comparing the average results between laboratories. Additionally, the variation observed within laboratory should be compared to ensure that laboratory does not have an effect either on the average result of the method or on the variability of the method. To assess reproducibility, conduct the same set of validation experiments within each laboratory and compare both the accuracy results and the precision results. If the differences are meaningful, analysis of variance (ANOVA) tests can be conducted to determine whether there is a statistically significant laboratory effect on the mean or on the variance of the method. For simplicity, the validation discussed within this chapter will not consider reproducibility and only one laboratory is considered.

Intermediate precision components typically include day, operator, and assay. Day is a random effect that captures random environmental changes that are not controlled and may have an effect on the assay result. Operator, usually modeled as a random effect, captures the variation in assay results due to change in personnel running the assay. Assay captures the variation one would expect from one complete run of the method to the next complete run of the method. Thus, assay captures the random variation due to slight perturbations in the sample preparation from assay run to assay run (when run by the same operator on the same day). Often, it is not possible for each operator to run more than one assay on each day; however, the validation experiments can be designed to estimate this variance component under a reasonable assumption.

ICH Q2A defines repeatability as the variability of the assay results “under the same operating conditions over a short interval of time. Repeatability is also termed intra-assay precision.” A reportable value of an assay is often the average of a specified number of replicate values, where the replicates are processed using the same sample preparation. Thus, it is not possible to obtain a true repeat of the assay’s reportable value. Repeatability can be modeled as the within-assay variability of the replicates; however, it should be noted that this precision is at the level of the replicate and not at the level of the reportable value. If the reportable value is an average of $K$ replicates the variance of the reportable values will be less than that of the replicates by a factor of $1/K$ (the standard deviation will be less by a factor of $(1/\sqrt{K})$.

Ruggedness is not mentioned in the ICH guidelines. The USP (5) defines ruggedness as “the degree of reproducibility of test results obtained by the analysis of the same samples under a variety of normal test conditions.” Thus, ruggedness also addresses the typical external changes in the execution of the method such as change in day, operator, or assay preparations. Ruggedness and intermediate precision are two sides of the same coin, but they measure different concepts. Intermediate precision is measured with a standard deviation or %RSD. Ruggedness is the lack of a factor effect. “Changing the instrument has no effect on the results” is a statement addressing ruggedness. Ruggedness is modeled as a fixed effect measuring the change in average results due to a change in instrument.

## Linearity

A method is “linear” or operating in the linear region if the measurement of the product parameter is proportional to the quantity being measured. Although there seems to be some debate regarding this issue, “linearity” does not refer to the linear range of a calibration curve used within the method or assay for calculation of the reportable value, but rather to the proportionality of the method’s reportable value to the quantity being measured. Here too, it is important that the validation experiments are balanced both with respect to the variance components and also over the range of method validation. Thus, the levels of expected input should be evenly spaced across the range of the method with an equal number of observations at each level.

### Robustness

Robustness refers to the effect on the assay response due to minor perturbations in sample preparations. In contrast to precision, robustness refers to the minor perturbations of factors internal to the assay (e.g., incubation time, temperature, or amount of reagent) whereas intermediate precision and ruggedness refer to factors necessary in the performance of the method that are external to the assay (day-to-day changes, operator changes, instrument changes). In method development, acceptable ranges for assay input parameters should be found where the assay response stays fairly constant. In these ranges the method is said to be robust. To quote for Q2B of the ICH Guidelines: "The evaluation of robustness should be considered during the development phase and depends on the type of procedure under study. It should show the reliability of an analysis with respect to deliberate variations in method parameters (2)."

## DESIGN OF EXPERIMENTS

### Terms and Definitions

#### Factors, Levels, and Treatments

Factors are the major categories of a design. The levels of each factor are the "values" or "versions" that the factor can assume. Treatments are the combinations of the factor levels that are assigned to the experimental units. In our validation experiment there are factors of "operator" and "day," each having two different levels. There are four treatments in the example validation matrix (refer to Table 1 in the Example Protocol).

## Fixed Effects, Random Effects, Main Effects, and Interactions

Factors are fixed if the "values" of the levels are given or do not vary. Factors are random if the value is, for pragmatic purposes, randomly selected from a population of all values. To illuminate in an example: if a small lab has two analysts that will always be running the assays then the factor "analyst" is considered fixed with two levels (one level for analyst 1, the other level for analyst 2); however, if the lab is large and/or there is a high turnover in analysts then for each run of an assay the "analyst" could be regarded as selected at random among a population of trained analysts.

In a statistical model, fixed effects have an influence on the mean value or average of the method's response while random effects have an influence on the variability of the method. Fixed effects are assessed in the context of accuracy. Random effects are assessed in the context of precision and become the intermediate precision components. In designing the validation design matrix the validation assays need to be balanced over both the fixed effects and the random effects. A mixed effects model (or design) occurs when both fixed effects and random effects are present (6).

Only one lot of well-characterized product or reference standard should be used in the validation. Product lot or batch is not a factor in method validation. Factors that are typically assessed are instrument or machine, analyst or operator, day, and other factors external to the method itself that come into play in the typical running of an assay.

Main effects of a factor are the differences in average response between the levels of the factor averaged over all other levels of the other factors. For example, if the factor machine has two levels, machine 1 and machine 2, the main effect, due to machine, would be the difference in average response from machine 1 and the average response from machine 2 (averaged over all other factors). An interaction occurs when the effect of one factor is dependent on the level of another factor. For example, if there is a significant increase in average response between the low and high levels of factor A when factor B is low and there is a significant decrease in

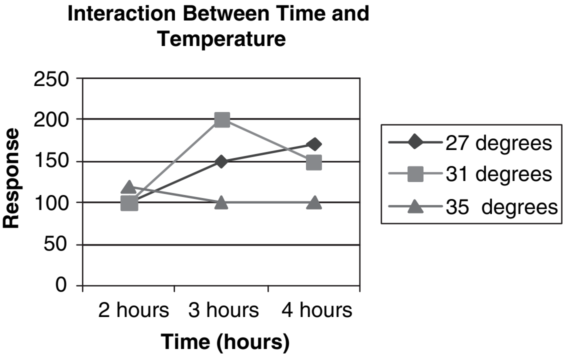


**Figure 1** Interaction between time and temperature. The response profile to time depends on the temperature.

average response between the low and high levels of factor A when factor B is high then there is an interaction between factor A and factor B. Figure 1 shows an interaction between time and temperature.

## Nested and Crossed Factors

Factor A is nested within factor B if the levels of factor A change with the level of factor B. For example, if batches or lots are tested by taking samples from the production line and then aliquots are taken from each sample, then the aliquots are nested within the samples which are nested within the batches (7). Factors C and D are crossed if all levels of factor C can occur with all levels of factor D. For example, in the sample protocol of section IIB, days and operators are crossed, as each of the two operators perform an assay on each of two the days.

## Aliasing and Confounding

If each level of one factor only occurs at a specific level of another factor, then the two factors are said to be confounded. When confounding or aliasing occurs, the data analysis cannot distinguish between the effects of the two confounded factors.

If a significant effect is apparent it is impossible to tell which factor is responsible for the effect. In designing experiments, the number of treatments or experimental trials can be reduced by intentionally aliasing one factor with another. For example, a main effect can be aliased with a high-order interaction, especially if it is a reasonable assumption that the interaction effect does not exist or is negligible.

## Experimental Designs

In this section, three categories of experimental design are considered for method validation experiments. An important quality of the design to be used is balance. Balance occurs when the levels each factor (either a fixed effects factor or a random effects variance component) are assigned the same number of experimental trials. Lack of balance can lead to erroneous statistical estimates of accuracy, precision, and linearity. Balance of design is one of the most important considerations in setting up the experimental trials. From a heuristic view this makes sense, we want an equivalent amount of information from each level of the factors.

### Full Factorial Designs

A full factorial design is one in which every level of each factor occurs for every combination of the levels of the other factors. If each of $n$ factor has two levels, frequently denoted "+" for one level and "–" for the other level, then the number of treatments is two raised to the $n$th power. Full factorial designs when each factor has the same number of levels are referred to as $k^n$ factorial designs, where $k$ is the number of levels and $n$ is the number of factors, as $k^n$ is the number of treatments needed. If factor A has three levels, factor B has two levels, and factor C has three levels then a full factorial design has $3 \times 2 \times 3 = 18$ treatments or experimental trials.

The treatments of a design (or the experimental trials) can be represented in a design matrix. Each row of the matrix represents a treatment; each column represents the level of each factor. A full factorial design for three factors, factor A,

**Table 4** Full Factorial Design for Three Factors, Factors A, B, and C, Each with Two Levels, Low (–) and High (+)

| Treatment | Factor A | Factor B | Factor C |
|---|---|---|---|
| 1 | – | – | – |
| 2 | + | – | – |
| 3 | – | + | – |
| 4 | + | + | – |
| 5 | – | – | + |
| 6 | + | – | + |
| 7 | – | + | + |
| 8 | + | + | + |

Every level of each factor occurs with every level of the other factors and combination of the other two factors.

factor B, and factor C, each with three levels, is shown in the design matrix presented in Table 4.

### Fractional Factorial Designs

Often a full factorial design requires too many resources and an experimental design with fewer treatments is required. Fractional factorials can reduce the number of experimental trials by aliasing one or more of the factors with other factors or interactions of factors. A fractional factorial of a full factorial design $2^n$ cuts the number of trials or treatments by1/2, 1/4, or $1/2^p$. When the number of experimental trials is reduced from $2^n$ by a factor of $1/2^p$. the design is referred to as a $2^{n-p}$ fractional factorial. The full factorial design ($2^3$ factorial design with eight treatments) presented in Table 4 can be reduced to a $2^{3-1}$ fractional factorial design with four experimental trials if factor C is aliased with the two-factor interaction of factors A and B. The resulting design is presented in Table 5. Note that the effect of factor C is confounded with the interaction of factors A and B. This can be observed in the design matrix by multiplying the levels of factors A and B to get the level of factor C (+ × + = +, + × – = –, and – × – = +).

The resolution of a design is a categorization of the aliasing in the design. For resolution III design, main effects are not aliased with each other but each main effect is aliased

**Table 5** Fractional Factorial Design for Three Factors, Factors A, B, and C, Each with Two Levels, Low (–) and High (+)

| Treatment | Factor A | Factor B | Factor C |
|---|---|---|---|
| 1 | – | – | + |
| 2 | + | + | + |
| 3 | – | + | – |
| 4 | + | – | – |

Factor C is aliased with the interaction of factors A and B (refer to text).

with two-factor interactions. In a resolution IV design, some main effects are aliased with three-factor interactions and two-factor interactions are aliased with each other (2–4). The design shown in Table 5 is resolution III, as main effect of factor C is aliased with the interaction between factors A and B. Also, A is aliased with B and C, and B is aliased with A and C. Resolution of design is discussed in more detail in most texts on experimental design including John (6), Box, Hunter, and Hunter (7), and Kuehl (8).

### Plackett-Burman Designs

Plackett-Burman designs are a group of orthogonal array designs of resolution III, where the number of experimental trials is a multiple of four, but is not a power of two. In most Plackett-Burman designs each factor has exactly two levels. These designs are useful as they require fewer resources. Placket-Burman designs are used to estimate main effects only. The reader is referred to the references (6–8) on experimental design for listings of Plackett-Burman designs.

## Designs for Quantitative Method Validation

Full factorial designs can be used in quantitative method validation. With very few factors this is a feasible design. A simple way to display the experiment necessary for the validation is to display the assay runs in a table or matrix. For instance, suppose a method is run on two different machines and the goal is to assess intermediate precision components. We have the random effects of operator and day and a fixed effect due to machine. If the factors operator and day each have two levels, the full factorial design is shown in Table 4 with factor A as operator, factor B as day, and factor C as machine. Eight assays will be run; the replicates of the assays will be used in estimating repeatability. All factors are crossed with each other.

The full factorial design may not be practical. Perhaps it is not possible for one operator to run two assays on one day. Then a fractional factorial design can be used. The design shown in Table 5 is resolution III design with four experimental trials. As machine is aliased with the interaction between operator and day, the main effect due to machine can be estimated only if we assume there is negligible interaction between operator and day. Often, interactions can be assumed to be negligible based on prior knowledge or the science that is known about the situation.

The design shown in Table 5 is also the design of the example protocol. In the example protocol the third factor is assay. The effect of assay is aliased with the interaction of operator and day, and assay is also nested within the interaction. Note that if the design of Table 4 were used, and two assays were run for each operator and day combination, then assay would still be nested within the interaction. Each time an assay is run, it is a different assay; hence, the number of levels of the factor assay is just the number of assays that are run. However, running two assays within each operator and day combination will give a better empirical estimate of inter-assay variability but assay effect cannot be estimated independently of the interaction. This is because assay takes on a different "value" each time the assay is run. (In contrast, in the previous example the third-factor machine is crossed with the factors operator and day; consequently, the machine effect can be estimated independently of operator and day effects.)

For quantitative methods, the ICH guidelines recommend validating a range that extends 20% below the lower end of the range and 20% above the upper end of the range (for content uniformity the range is extended to 70–130% of the SOP range). For the linearity portion of the method, National Committee for Clinical Laboratory Standards [NCCLS (9)] recommends at least five levels evenly spaced across the range to be validated. For example in validating a method that measures concentration of an analyte, the sample being tested would be diluted out to (or concentrated to) five expected concentrations evenly spaced between 20% below the lower limit of the method's stated range and 20% above the upper limit of the method's stated range. In an assay that can measure multiple concentrations simultaneously, each assay would test these five diluted samples in $K$-replicates (within-assay replicates; the assay's reportable value could be the average of these $K$ replicates). If this procedure is repeated using $I$ operators over $J$ days then the method characteristics of accuracy, precision, and linearity can be assessed with one set of $I \times J$ assays. In essence, the factor concentration is crossed with the other factors as each of the five concentrations is run in all of the assays. In the example protocol, eight levels of mass are used.

## STATISTICAL ANALYSES

The experimental design selected, as well as the type of factors in the design, dictates the statistical model to be used for data analysis. As mentioned previously, fixed effects influence the mean value of a response, while random effects influence the variance. In this validation, the model has at least one fixed effect of the overall average response and the intermediate precision components are random effects. When a statistical model has both fixed effects and random effects it is called a mixed effects model.

The statistical software package SAS® (SAS Institute Inc., Cary, North Carolina, U.S.A.) has a module [PROC MIXED (10)], which analyzes mixed effects models and provides estimates for both the fixed (shifts on the mean) and random effects (variance components estimates).

Analysis of accuracy and precision can be accomplished using the same statistical model with a slightly different response variable. For accuracy, the response variable (here this is the replicate value and differs from the reportable value of the assay as run outside of the validation) is observed mass divided by expected mass as a percentage. For precision the response variable is observed mass. Both the accuracy and the precision results are reported for each expected mass. The linearity assessment uses the entire data set over all expected masses. Consequently, the root mean square error (RMSE) from the linearity analysis can be viewed as measurement of overall repeatability.

ANOVA can be used to test whether the fixed effects are significant. For example, if there is a factor machine with two levels (representing two machines), an ANOVA can be used to estimate machine effect. If the difference between the average responses for each machine is not meaningfully different, and the variance components within each machine are similar, then the variance components can be analyzed averaging over the machines. If there is a meaningful difference between the two observed machine averages, then an *F*-test can be used to test whether machine effect is significant.

Accuracy is estimated from the fixed effects components of the model. If the overall mean is the only fixed effect, then accuracy is reported as the estimate of the overall mean accuracy with a 95% confidence interval. As the standard error will be calculated from a variance components estimate including intermediate precision components and repeatability, the degrees of freedom can be calculated using Satterthwaite's approximation (6). The software program SAS has a procedure for mixed model analysis (PROC MIXED); PROC MIXED has an option to use Satterthwaite's degrees of freedom in calculating the confidence interval for the mean accuracy. An example program and output is shown later for the example protocol.

Precision components are estimated using the random effects of the model. There are a number of ways to estimate the precision components (or variance components). Variance components are estimates of variances (variance = standard deviation squared) and consequently are restricted to being nonnegative. One approach is to conduct an ANOVA and solve for the variance components using the theoretical composition of the expected mean squares (called method of moments estimates or ANOVA estimates). The ANOVA method is shown in many textbooks, including Searle et al. (11) which provides an in-depth discussion of variance components estimation. Another method is to use maximum likelihood estimation (MLE). Both of these approaches have drawbacks. ANOVA estimates can be negative as they are a linear combination of sums of squares and there is nothing inherent in restricting the resulting estimates to be nonnegative. Maximum likelihood estimates can be biased as the estimators do not factor in the number of fixed effects being estimated in the model. A third approach is to use maximum likelihood and restrict the domain of the variance estimates to nonnegative values by "estimating variance components based on residuals calculated after fitting by ordinary least squares just the fixed effects part of the model" (11); this is known as restricted maximum likelihood estimation (REML). If the design is balanced (i.e., there are the same number of assays run per each day, per each operator, and so on) and the ANOVA estimates are nonnegative, then theoretically the REML estimates and the ANOVA estimates agree. SAS's PROC MIXED has options for all of these methods of estimation. The example protocol analysis shows use of REML. For any of these methods, it is very important that the original design of experiments is balanced to get statistically sound estimates of variance components.

The basic idea behind variance components estimation is that the variability of the data is parsed out to the random effects of the model. For example, operator-to-operator variability can be assessed roughly by calculating the standard deviation of the operator-specific averages. Thus, it is clear why balance is very important; if the data are not balanced then we are comparing averages from differing number of values, thus the average with the smaller sample size is more variable than that with the larger sample size (i.e., we are mixing apples with oranges).

Total precision is defined as the sum of the precision components. This summation takes place on the variance scale (variance = standard deviation squared). In theory, variances from independent terms in a model add up to total variance. See the precision analysis in the validation analysis example of the next section.

Linearity is analyzed using least squares regression of the observed responses against the expected responses, or the regression of the observed responses against the amount of analyte in the sample. The linearity here is of the reportable value versus the amount of analyte in the sample. Note, in the example protocol the responses are the replicate values for each assay at each mass; the average of three within-assay replicates is the reportable value (at each mass) of SOP 123 (in release testing, only one reportable value, the average of three replicates, is reported as the mass of the sample is unknown). If the entire range is assayed in each assay (as in the example validation), then linearity can be assessed within each assay or over all assays. If linearity is assessed using all the data then the RMSE can be used as a measure of overall precision as all the assays over all days and operators are represented. If linearity is assessed within each assay then the RMSE is another measure of repeatability (measurement of error variability within assay).

When there is more than one observation (whether it be replicate values or more than one reportable value over a number of assays) at each level of the analyte, then a lack-of-fit analysis can be conducted. This analysis tests whether the average response at each level of the analyte is a significantly better model for average assay response than the linear model. A significant lack-of-fit can exist even with a high correlation coefficient (or high coefficient of determination) and the maximum deviation of response from the predicted value of the line should be assessed for practical significance.

## EXAMPLE VALIDATION DATA ANALYSIS

The data are collected and recorded in tables (Tables 6A–D for hypothetical data). The validation characteristics of accuracy, precision, and linearity are analyzed from the data for this example validation.

### Model for Accuracy and Precision

The statistical model used to analyze both accuracy and precision is a mixed effects model. Separately for each expected

**Table 6A** Data from Assay A for Example Validation

| Operator | Day | Replicate | Expected | Observed |
|---|---|---|---|---|
| 1 | 1 | 1 | 40 | 38.7398 |
| 1 | 1 | 2 | 40 | 38.6092 |
| 1 | 1 | 3 | 40 | 39.0990 |
| 1 | 1 | 1 | 50 | 49.4436 |
| 1 | 1 | 2 | 50 | 48.7904 |
| 1 | 1 | 3 | 50 | 48.6924 |
| 1 | 1 | 1 | 70 | 70.6760 |
| 1 | 1 | 2 | 70 | 69.5980 |
| 1 | 1 | 3 | 70 | 70.2180 |
| 1 | 1 | 1 | 90 | 88.7400 |
| 1 | 1 | 2 | 90 | 87.7280 |
| 1 | 1 | 3 | 90 | 88.4780 |
| 1 | 1 | 1 | 110 | 106.3140 |
| 1 | 1 | 2 | 110 | 105.2700 |
| 1 | 1 | 3 | 110 | 106.3780 |
| 1 | 1 | 1 | 130 | 123.5940 |
| 1 | 1 | 2 | 130 | 122.5820 |
| 1 | 1 | 3 | 130 | 122.1900 |
| 1 | 1 | 1 | 150 | 138.5220 |
| 1 | 1 | 2 | 150 | 138.9480 |
| 1 | 1 | 3 | 150 | 138.9480 |
| 1 | 1 | 1 | 180 | 172.9140 |
| 1 | 1 | 2 | 180 | 172.2600 |
| 1 | 1 | 3 | 180 | 173.7300 |

mass of 40 $\mu$g, 50 $\mu$g, 70 $\mu$g, 90 $\mu$g, 110 $\mu$g, 130 $\mu$g, 150 $\mu$g, and 180 $\mu$g:

$$\mathrm{Y}_{ijk} = \mu + \mathrm{Op}_i + \mathrm{Day}_j + \mathrm{Assay}\ (\mathrm{Op} \times \mathrm{Day})_{ij} + \varepsilon_{ijk} \qquad (1)$$

Here, *Y* is the response variable. For accuracy, *Y* is observed mass divided by expected mass expressed as a percent, % accuracy (note that this is measuring percent recovery against an expected value). For precision, *Y* is observed mass. The term $\mu$ is a fixed effect representing the population overall average mass for each expected mass. The term Op is the random component added by operator *i*; Day is the random component added by day *j*; and Assay is nested in the interaction of operator and day. The term ε is the random error of replicate *k* of the assay performed by operator *i* on day *j*.

**Table 6B** Data from Assay B for Example Validation

| Operator | Day | Replicate | Expected | Observed |
|---|---|---|---|---|
| 1 | 2 | 1 | 40 | 38.0584 |
| 1 | 2 | 2 | 40 | 38.1000 |
| 1 | 2 | 3 | 40 | 38.0450 |
| 1 | 2 | 1 | 50 | 47.6778 |
| 1 | 2 | 2 | 50 | 47.7412 |
| 1 | 2 | 3 | 50 | 48.1222 |
| 1 | 2 | 1 | 70 | 67.4260 |
| 1 | 2 | 2 | 70 | 66.9500 |
| 1 | 2 | 3 | 70 | 66.9500 |
| 1 | 2 | 1 | 90 | 86.7940 |
| 1 | 2 | 2 | 90 | 85.6180 |
| 1 | 2 | 3 | 90 | 86.1580 |
| 1 | 2 | 1 | 110 | 103.4300 |
| 1 | 2 | 2 | 110 | 101.6200 |
| 1 | 2 | 3 | 110 | 102.0960 |
| 1 | 2 | 1 | 130 | 118.4780 |
| 1 | 2 | 2 | 130 | 117.4940 |
| 1 | 2 | 3 | 130 | 118.2260 |
| 1 | 2 | 1 | 150 | 134.7020 |
| 1 | 2 | 2 | 150 | 132.1940 |
| 1 | 2 | 3 | 150 | 132.5440 |
| 1 | 2 | 1 | 180 | 179.2500 |
| 1 | 2 | 2 | 180 | 172.1000 |
| 1 | 2 | 3 | 180 | 174.5000 |

Average accuracy is estimated by the overall observed accuracy. The parameter being estimated is $\mu$. The standard error of this estimate is a function of the variance components for operator, day, assay, and repeatability.

SAS PROC MIXED is used to analyze the data. For accuracy, the SAS code used is:

```
*SAS code for Accuracy;
PROC MIXED DATA = example CL;
CLASS oper day assay;
MODEL acc = / solution cl ddfm = satterth;
RANDOM oper day assay;
BY exp;
RUN
```

**Table 6C** Data from Assay C for Example Validation

| Operator | Day | Replicate | Expected | Observed |
|---|---|---|---|---|
| 2 | 1 | 1 | 40 | 41.7500 |
| 2 | 1 | 2 | 40 | 43.2800 |
| 2 | 1 | 3 | 40 | 43.5100 |
| 2 | 1 | 1 | 50 | 48.1538 |
| 2 | 1 | 2 | 50 | 47.9650 |
| 2 | 1 | 3 | 50 | 48.3110 |
| 2 | 1 | 1 | 70 | 67.9920 |
| 2 | 1 | 2 | 70 | 67.8340 |
| 2 | 1 | 3 | 70 | 68.0220 |
| 2 | 1 | 1 | 90 | 88.6140 |
| 2 | 1 | 2 | 90 | 87.1980 |
| 2 | 1 | 3 | 90 | 86.9780 |
| 2 | 1 | 1 | 110 | 104.1140 |
| 2 | 1 | 2 | 110 | 103.3580 |
| 2 | 1 | 3 | 110 | 103.6100 |
| 2 | 1 | 1 | 130 | 119.9580 |
| 2 | 1 | 2 | 130 | 119.4560 |
| 2 | 1 | 3 | 130 | 119.2360 |
| 2 | 1 | 1 | 150 | 136.9660 |
| 2 | 1 | 2 | 150 | 133.6340 |
| 2 | 1 | 3 | 150 | 134.3260 |
| 2 | 1 | 1 | 180 | 179.1500 |
| 2 | 1 | 2 | 180 | 178.2500 |
| 2 | 1 | 3 | 180 | 178.9100 |

In the code, “oper” reads in the level of operator, “day” reads in the level of day, “assay” reads in the assay, “acc” is *Y* for accuracy and is defined by the ratio of observed mass to expected mass (“exp”). Note that the “BY exp” the procedure is run separately for each level of the expected mass using “exp” as the variable name. In this manner, the precision of the accuracy and the variance components are estimated independently at each level of the expected mass. Thus, we obtain “picture” of the assay’s performance characteristics across the operational range. As assay is nested in the interaction between operator and day, the same analysis can be coded used “oper* day” in place of “assay” in the previuos code. The “CL” and “cl” in the procedure line and model statement are

**Table 6D** Data from Assay D for Example Validation

| Operator | Day | Replicate | Expected | Observed |
|---|---|---|---|---|
| 2 | 2 | 1 | 40 | 41.4500 |
| 2 | 2 | 2 | 40 | 41.7500 |
| 2 | 2 | 3 | 40 | 41.6000 |
| 2 | 2 | 1 | 50 | 48.5228 |
| 2 | 2 | 2 | 50 | 49.1456 |
| 2 | 2 | 3 | 50 | 49.3636 |
| 2 | 2 | 1 | 70 | 68.8860 |
| 2 | 2 | 2 | 70 | 70.0380 |
| 2 | 2 | 3 | 70 | 70.1620 |
| 2 | 2 | 1 | 90 | 86.9140 |
| 2 | 2 | 2 | 90 | 87.9740 |
| 2 | 2 | 3 | 90 | 88.5340 |
| 2 | 2 | 1 | 110 | 104.8500 |
| 2 | 2 | 2 | 110 | 104.5680 |
| 2 | 2 | 3 | 110 | 104.9740 |
| 2 | 2 | 1 | 130 | 120.3240 |
| 2 | 2 | 2 | 130 | 120.0440 |
| 2 | 2 | 3 | 130 | 121.3820 |
| 2 | 2 | 1 | 150 | 136.1720 |
| 2 | 2 | 2 | 150 | 136.7960 |
| 2 | 2 | 3 | 150 | 136.6380 |
| 2 | 2 | 1 | 180 | 173.2800 |
| 2 | 2 | 2 | 180 | 183.2000 |
| 2 | 2 | 3 | 180 | 175.3200 |

calls for confidence limits. In the model statement "ddfm = satterth" calls for Satterthwaite's degrees of freedom (see Ref. 6 for a discussion of Satterthwaite's approximation) to be used in calculating the confidence limits.

Model (1) is used for precision analysis. For precision, $Y$ is the observed mass (rather than observed mass divided by expected mass). Each random effects component (operator, day, assay, and error or repeatability) contributes to the total variance of $Y$. The total variance of $Y$ is defined as the sum of the variance components. If $\sigma^2_{Oper}$, $\sigma^2_{Day}$, $\sigma^2_{Assay}$, and $\sigma_\varepsilon^2$ are the variance components for operator, day, assay, and repeatability, respectively, then total variance is $\sigma^2_{Oper}+ \sigma^2_{Day}+ \sigma^2_{Assay}+ \sigma^2_{\varepsilon}$.

For precision, the SAS code used is shown next. Note that the only change is the *Y* value; for precision, *Y* is observed mass, called "obt" in the code.

```
*SAS code for Precision;
PROC MIXED DATA = example CL;
CLASS oper day assay;
MODEL obt = / solution cl ddfm = satterth;
RANDOM oper day assay;
BY exp;
RUN
```

This program will give variance estimates for each of the precision components along with two-sided 95% confidence intervals for the population variance component for each expected mass. SAS PROC MIXED will provide ANOVA estimates, maximum likelihood estimates, and REML estimates; the default estimation, used here, is REML.

Method of moments estimates (also known as ANOVA estimates) can be calculated directly from the raw data as long as the design is balanced. The reader is referred to Searle et al. (11) for a thorough but rather technical presentation of variance components analysis. The equations that follow show the ANOVA estimates for the validation example. First, a two-factor with interaction ANOVA table is computed (Table 7). Then the observed mean squares are equated to the expected mean squares and solved for the variance components (Table 8 and the equations that follow).

The expected mean squares are equated to the observed mean squares (those from the data) and solved for the variance components. Thus, $\hat{\sigma}^2_\varepsilon = \text{MSE}$ is the variance estimate for repeatability, $\hat{\sigma}^2_{Assay} = \frac{\text{MSA} - \text{MSE}}{m}$ is the variance estimate for assay, $\hat{\sigma}^2_{Day} = \frac{\text{MSD} - \text{MSA}}{a \cdot m}$ is the variance estimate for day, and $\hat{\sigma}^2_{Oper} = \frac{\text{MSO} - \text{MSA}}{b \cdot m}$ is the variance estimate for operator. To express as standard deviations, take the square root of each. The ANOVA estimates are applicable when the design is balanced and there are no missing data values; using this method with missing data or an unbalanced design will result in misleading erroneous variance estimates.

**Table 7** Analysis of Variance (ANOVA), Sum of Squares and Mean Square Formulae for a Two-Factor Mixed Model with Interaction for the Validation Matrix of Tables 4 or 5 to be Used in ANOVA Variance Component Estimation

| Source | *df* | Sum of squares | Mean square |
|---|---|---|---|
| Operator | $a-1$ | $\text{SSO} = \sum \text{b} \cdot \text{m}(\bar{\bar{y}}_{i..} - \bar{\bar{\bar{y}}}_{...})^2$ | $\text{MSO} = \frac{\text{SSO}}{df\text{O}}$ |
| Day | $b-1$ | $\text{SSD} = \sum \text{a} \cdot \text{m}(\bar{\bar{y}}_{.j.} - \bar{\bar{\bar{y}}}_{...})^2$ | $\text{MSD} = \frac{\text{SSD}}{df\text{D}}$ |
| Assay = operator × day | $(a-1) \times (b-1)$ | $\text{SSA} = \sum\sum \text{m}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{\bar{\bar{y}}}_{...})^2$ | $\text{MSA} = \frac{\text{SSA}}{df\text{A}}$ |
| Repeatability | $ab(m-1)$ | $\text{SSE} = \sum\sum\sum(y_{ijk} - \bar{\bar{\bar{y}}}_{...})^2$ | $\text{MSE} = \frac{\text{SSE}}{df\text{E}}$ |
| Total | $abm-1$ | $\text{SST} = \sum\sum\sum(y_{ijk} - \bar{\bar{\bar{y}}}_{...})^2$ | |

Here, a bar over the *Y* indicates averaging over the missing subscript(s); the subscript "*i*" is an index for operator, the subscript "*j*" is an index for day, assay is indexed by the level of "*i*" and "*j*," and "*k*" is an index for the replicate within each assay. In the example, the grand average is $\bar{\bar{\bar{y}}}_{...}$, the average for operator "*i*" is $\bar{y}_{i.} = \sum_j \sum_k y_{ijk} / 6$, the average of the six replicates produced by operator "*i*" for $i = 1, 2$, and the average for assay "*i, j*" is the average of the three replicates of operator "*i*" on day "*j*." The design must be balanced for these formulas to apply.
*Abbreviations*: *a*, number of operators; *b*, number of days; *m*, number of replicates; *df*, degrees of freedom.
*Source*: From Ref. 11.

## Linearity Analysis

For the example validation, linearity will be analyzed within each assay and across all four assays. For the second analysis, as mentioned before, the RMSE can be used as a measure of

**Table 8** Expected Mean Squares for the Two-Factor with Interaction Analysis of Variance Table Shown in Table 7

| Source | Mean square | Expected mean square |
|---|---|---|
| Operator | $\text{MSO} = \frac{\text{SSO}}{df\text{O}}$ | $\sigma^2_\varepsilon + m \cdot \sigma^2_{Assay} + b \cdot m \cdot \sigma^2_{Oper}$ |
| Day | $\text{MSD} = \frac{\text{SSD}}{df\text{D}}$ | $\sigma^2_\varepsilon + m \cdot \sigma^2_{Assay} + a \cdot m \cdot \sigma^2_{Day}$ |
| Assay = operator × day | $\text{MSA} = \frac{\text{SSA}}{df\text{A}}$ | $\sigma^2_\varepsilon + m \cdot \sigma^2_{Assay}$ |
| Repeatability | $\text{MSE} = \frac{\text{SSE}}{df\text{E}}$ | $\sigma^2_\varepsilon$ |

overall precision. For the first analysis, the RMSE is a measure of repeatability across all masses within each assay. The model for linearity is:

$$\text{Observed mass} = \alpha + \beta \cdot \text{Expected mass} + \varepsilon \tag{2}$$

Here, $\alpha$ is the population intercept and $\beta$ is the population slope of observed mass regressed on expected mass. The error term, $\varepsilon$, is the residual added to the population line to obtain the observed mass. Least squares estimation (traditional regression estimation) will yield estimates of $\alpha$ and $\beta$, as well as an estimate for the standard deviation of ε (RMSE). These estimates form the sample regression line for predicting observed mass (predicted mass or $\hat{Y}$):

$$\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot \text{Expected mass} \tag{3}$$

The RMSE is standard deviation of the residuals (residual = observed mass – predicted mass).

To test for lack-of-fit, a classification variable is added to the model representing the expected mass group:

$$\text{Observed mass} = \alpha + \text{Mass class} + \beta \cdot \text{Expected Mass} + \text{e} \tag{4}$$

Note that this analysis can only be performed if there are replicate values for each level of expected mass (mass class in the previous model). The model testing lack-of-fit is shown next. An ANOVA *F*-test is used to test whether the classification variable is significant, indicating significant (but not necessarily a meaningful) lack-of-fit to the linear model. In essence, this test asks the question "Do the averages at each expected mass fit the data better than the linear regression model?" If the method is very precise it will not take a very large deviation of observed mass from predicted mass to result in significant lack-of-fit. As a result, the deviations themselves should be analyzed for a meaningful lack-of-fit.

The two programs (SAS PROC GLM) for linearity are shown next for analyzing linearity within each of the four assays. The first gives the least squares regression estimates and the regression statistics; "obt" is the observed mass and "exp" is the expected mass for each of the eight masses assayed. The second program tests for lack-of-fit to the linear model; a new classification variable "expclass" has eight settings corresponding to the eight expected masses (40 $\mu$g, 50 $\mu$g, 70 $\mu$g, 90 $\mu$g, 110 $\mu$g, 130 $\mu$g, 150 $\mu$g, and 180 $\mu$g).

```
* To give least squares estimates for linear regression;
PROC GLM data = example;
Title 'obt = obtained mass, exp = expected mass';
MODEL obt = exp;
BY assay;
OUTPUT out = robt p = yhat r = res;
RUN;
*To test lack-of-fit to linear model;
PROC GLM data = example2;
Title 'Lack of fit analysis';
CLASS expclass ;
MODEL obt = exp expclass/solution;
BY assay;
RUN
```

The output statement in the first program creates a file with the data, the predicted masses by the linear regression model ($\hat{Y}$), and the residuals of the observed values minus the predicted values for each replicate at each expected mass. These deviations can be assessed for meaningful lack-of-fit.

## Example Protocol Analysis Results

This section shows the analysis results for the data of the validation example (Tables 6A–D). The results of the statistical analysis should be reported in a manner that is in accordance with company quality standards. For example, documentation of data auditing of the data listing from SAS against the raw data, a complete referencing to the electronic storage locations of the data, the SAS program, the program output, and a hardcopy write up of the statistical analysis approach should all be maintained with the validation report. For some companies, this usually results in a statistical report that is on the order of 150 to 200 pages in length (the entire SAS output is often included with the report). The tables and graphs reported as well as some examples of SAS output will be presented here. The report should enable another statistical analyst to reproduce the analysis at a later date.

## Accuracy

The model used for accuracy (1) is shown in the previous section (Example Protocol section IIB); recall that accuracy for each data point is defined as Accuracy = (Observed mass/Expected mass)×100%. The assay-specific averages for accuracy, the across-assay average, standard error of the across-assay average, Satterthwaite degrees of freedom, and 95% confidence intervals are shown in Table 9. Note that because Satterthwaite degrees of freedom are a linear combination of sums of squares, they can assume noninteger values.

An Excel chart showing the average accuracy and 95% confidence limits for each of the expected masses is shown next (Fig. 2). The confidence intervals tend to be wide if one (or more) of the intermediate precision components has a large contribution to variability.

The acceptance criteria for accuracy is that each observed average accuracy must be within the interval (90%, 100%) for the expected masses within SOP 123's stated range of 50 to 150 $\mu$g. As all the observed averages are within the interval (90%,100%), the acceptance criteria for accuracy have been met.

## Precision

The statistical model for precision is shown in equation (1). For precision, $Y$ is defined as observed mass. For each of the eight expected masses the precision components are reported as variances for each component. The total variance is the sum of the component variances. The component percent of total is calculated on the variance scale as: (Variance component estimate/Total variance)×100%. The square root of each variance estimate is taken and reported as the standard deviation. Confidence intervals were generated by SAS and reported on the variance scale; square roots of the confidence limits were calculated and reported as the confidence interval for the component standard deviation. The coefficient of variation (%CV) is calculated as: (Observed component std dev/Observed average)×100%. Table 10 shows the precision results for the example validation. The acceptance criteria for

**Table 9** Accuracy Results for the Example Validation Data

| Expected protein mass | Assay accuracy averages (%) | | | | Average accuracy | Standard | | 95% Confidence interval on accuracy | |
|---|---|---|---|---|---|---|---|---|---|
| ($\mu$g) | A | B | C | D | (%) | error (%) | *df* | Lower (%) | Upper (%) |
| 40 | 97.04 | 95.17 | 107.12 | 104.00 | 100.83 | 4.875 | 1.13 | 53.12 | 148.54 |
| 50 | 97.95 | 95.69 | 96.29 | 98.02 | 96.99 | 0.589 | 3 | 95.11 | 98.86 |
| 70 | 100.23 | 95.87 | 97.07 | 99.56 | 98.18 | 1.029 | 3 | 94.91 | 101.46 |
| 90 | 98.13 | 95.77 | 97.33 | 97.56 | 97.20 | 0.532 | 1 | 90.45 | 103.94 |
| 110 | 96.35 | 93.07 | 94.27 | 95.27 | 94.74 | 0.700 | 3 | 92.51 | 96.97 |
| 130 | 94.45 | 90.82 | 91.96 | 92.76 | 92.50 | 0.763 | 3 | 90.07 | 94.93 |
| 150 | 92.54 | 88.76 | 89.98 | 91.02 | 90.58 | 0.800 | 3 | 88.03 | 93.12 |
| 180 | 96.09 | 97.38 | 99.32 | 98.48 | 97.82 | 1.081 | 1 | 84.11 | 111.53 |

*Abbreviation*: df, degrees of freedom.

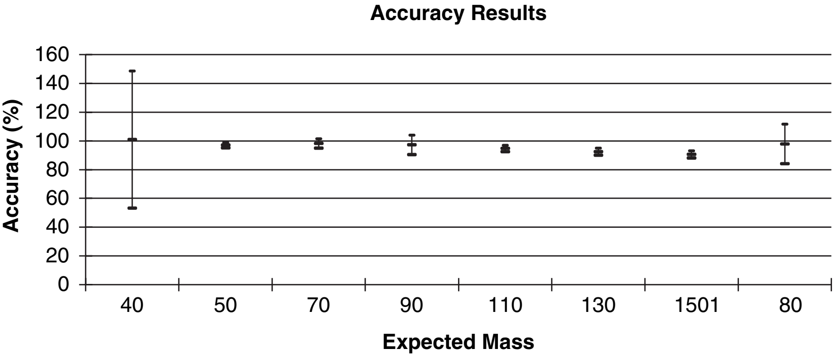


**Figure 2** Excel chart showing average accuracy with 95% confidence limits.

precision were that %CV must be at most 10%. The acceptance criteria were met.

### Linearity

The models used for analysis of linearity were shown in (2) and (4) before. Table 11 presents the summary statistics for the assay-specific regressions. Figures 3 to 6 show the assay scatter plots with regression lines. Table 12 shows the regression results for the lack-of-fit analysis; as each of the assays displayed a significant lack-of-fit, the maximum percent deviation is calculated as the maximum of: [(Observed Mass − Predicted mass)/Predicted mass] × 100%. In addition, the number of percent deviations over 10% is calculated for each expected mass. The acceptance criteria for linearity were that each and every coefficient of determination ($r^2$) must be at least 0.975. (For this analysis, the simple $r^2$ was calculated.) The acceptance criteria were met.

### Across-Assay Linearity

An alternative approach to linearity analysis is to analyze the linearity of all the data. The SAS code for the overall analysis is the same as shown before except that the line "By Assay"; is deleted from the code. Table 13 shows the summary statistics for the least squares regression, Figure 7 shows a scatter plot

(*Text continues on page 44.*)

**Table 10** Precision Results for the Example Validation Data

| Expected protein mass ($\mu$g) | Source | Standard deviation ($\mu$g) | Variance ($\mu$g$^2$) | Percent (%) | 95% Confidence interval on σ Lower | 95% Confidence interval on σ Upper | Estimated mean | CV (%) |
|---|---|---|---|---|---|---|---|---|
| 40 | Oper | 0.68 | 0.4568 | 5.85 | 0.29 | 40.29 | . | . |
| 40 | Day | 2.67 | 7.1090 | 91.03 | 1.19 | 88.39 | . | . |
| 40 | Assay | 0.00 | 0.0000 | 0.00 | . | . | . | . |
| 40 | Residual | 0.49 | 0.2436 | 3.12 | 0.34 | 0.90 | . | . |
| 40 | *Total* | 2.79 | 7.8094 | 100.00 | . | . | 40.33 | 6.93 |
| 50 | Oper | 0.00 | 0.0000 | 0.00 | . | . | . | . |
| 50 | Day | 0.00 | 0.0000 | 0.00 | . | . | . | . |
| 50 | Assay | 0.56 | 0.3097 | 73.57 | 0.30 | 2.72 | . | . |
| 50 | Residual | 0.33 | 0.1112 | 26.43 | 0.23 | 0.64 | . | . |
| 50 | *Total* | 0.65 | 0.4209 | 100.00 | . | . | 48.49 | 1.34 |
| 70 | Oper | 0.00 | 0.0000 | 0.00 | . | . | . | . |
| 70 | Day | 0.00 | 0.0000 | 0.00 | . | . | . | . |
| 70 | Assay | 1.42 | 2.0025 | 90.17 | 0.79 | 5.70 | . | . |
| 70 | Residual | 0.47 | 0.2184 | 9.83 | 0.32 | 0.90 | . | . |
| 70 | *Total* | 1.49 | 2.2209 | 100.00 | . | . | 68.73 | 2.17 |
| 90 | Oper | 0.26 | 0.0667 | 5.57 | 0.15 | $4.58E+101$ | . | . |
| 90 | Day | 0.00 | 0.0000 | 0.00 | . | . | . | . |
| 90 | Assay | 0.78 | 0.6091 | 50.85 | 0.36 | 14.51 | . | . |
| 90 | Residual | 0.72 | 0.5219 | 43.58 | 0.49 | 1.38 | . | . |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 90 | *Total* | 1.09 | 1.1977 | 100.00 | . | . | 87.48 | 1.25 |
| 110 | Oper | 0.00 | 0.0000 | 0.00 | . | . | . | . |
| 110 | Day | 0.00 | 0.0000 | 0.00 | . | . | . | . |
| 110 | Assay | 1.50 | 2.2487 | 86.04 | 0.83 | 6.28 | . | . |
| 110 | Residual | 0.60 | 0.3647 | 13.96 | 0.41 | 1.16 | . | . |
| 110 | *Total* | 1.62 | 2.6134 | 100.00 | . | . | 104.22 | 1.55 |
| 130 | Oper | 0.00 | 0.0000 | 0.00 | . | . | . | . |
| 130 | Day | 0.00 | 0.0000 | 0.00 | . | . | . | . |
| 130 | Assay | 1.95 | 3.8198 | 91.49 | 1.09 | 7.78 | . | . |
| 130 | Residual | 0.60 | 0.3553 | 8.51 | 0.40 | 1.14 | . | . |
| 130 | *Total* | 2.04 | 4.1751 | 100.00 | . | . | 120.25 | 1.70 |
| 150 | Oper | 0.00 | 0.0000 | 0.00 | . | . | . | . |
| 150 | Day | 0.00 | 0.0000 | 0.00 | . | . | . | . |
| 150 | Assay | 2.31 | 5.3345 | 80.70 | 1.27 | 10.26 | . | . |
| 150 | Residual | 1.13 | 1.2756 | 19.30 | 0.76 | 2.16 | . | . |
| 150 | *Total* | 2.57 | 6.6101 | 100.00 | . | . | 135.87 | 1.89 |
| 180 | Oper | 0.00 | 0.0000 | 0.00 | . | . | . | . |
| 180 | Day | 2.45 | 6.0013 | 38.88 | 0.98 | 593.64 | . | . |
| 180 | Assay | 0.00 | 0.0000 | 0.00 | . | . | . | . |
| 180 | Residual | 3.07 | 9.4323 | 61.12 | 2.15 | 5.39 | . | . |
| 180 | *Total* | 3.93 | 15.4340 | 100.00 | . | . | 176.07 | 2.23 |

*Abbreviation*: CV, coefficient of variance.

**Table 11** Linearity Summary Statistics for the Example Validation Data

| Assay | Correlation coefficient | Coefficient of determination | Intercept | Slope | Root mean square error | Residual sum of squares |
|---|---|---|---|---|---|---|
| A | 0.9989 | 0.9978 | 2.88 | 0.93 | 2.11 | 97.52 |
| B | 0.9956 | 0.9912 | 0.62 | 0.93 | 4.22 | 392.15 |
| C | 0.9946 | 0.9893 | 2.26 | 0.93 | 4.67 | 480.07 |
| D | 0.9960 | 0.9919 | 2.93 | 0.93 | 4.04 | 358.50 |

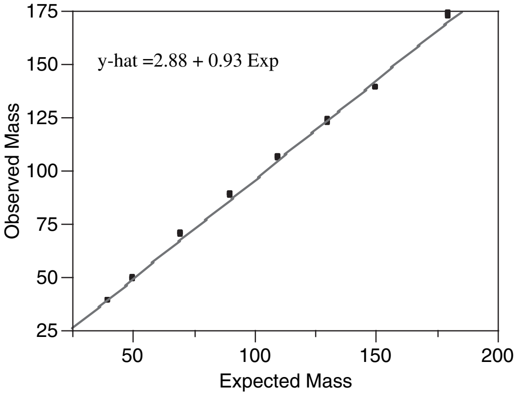


**Figure 3** Scatter plot of Assay A data with least-squares line.

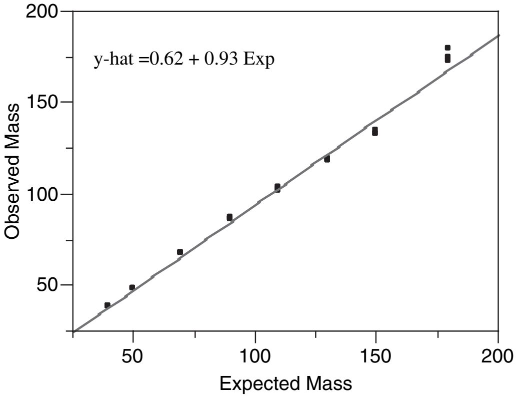


**Figure 4** Scatter plot of Assay B data with least-squares line.

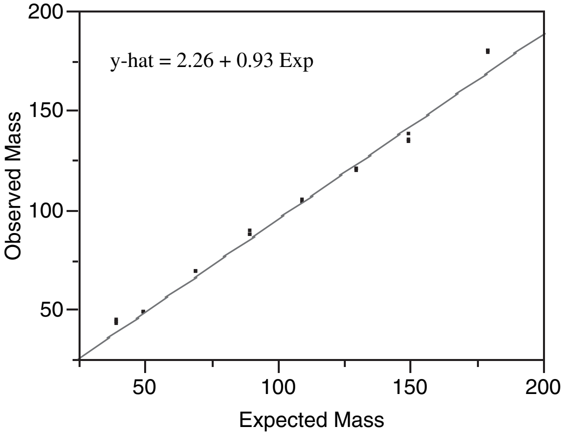


**Figure 5** Scatter plot of Assay C data with least-squares line.

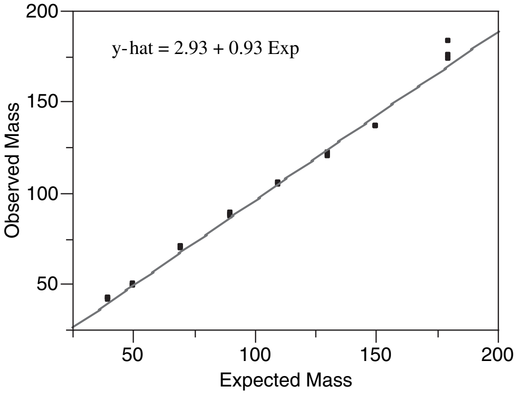


**Figure 6** Scatter plot of Assay D data with least-squares line.

**Table 12** Summary of Lack-of-Fit Analysis for Linearity of Example Validation Data

| Assay | Lack-of-fit $P$-value | Maximum deviation (%) | Level of maximum deviation ($\mu$g) |
|---|---|---|---|
| A | <0.0001 | 3.813 | 70 |
| B | <0.0001 | 6.611 | 180 |
| C | <0.0001 | 9.880 | 40 |
| D | <0.0001 | 7.381 | 180 |

**Table 13** Linearity Results for the Example Validation Data

| Correlation coefficient | Coefficient of determination | Intercept | Slope | Root mean square error | Residual sum of squares |
|---|---|---|---|---|---|
| 0.9960 | 0.9920 | 2.17 | 0.93 | 3.89 | 1420.63 |

All assays are used in the linear regression analysis; see Figure 7.

of the data with the fitted regression line and Table 14 summarizes the lack-of-fit analysis. Here, the simple $r^2$ (coefficient of determination) is calculated.

For single replicate observations, the highest absolute departure from the line occurs at 180 $\mu$g in assay D (13.31 $\mu$g difference from the line on the $y$-axis); the highest percentage difference occurs at 40 $\mu$g in assay C (10.31% difference). Using the average over all 12 observations at each mass, the largest absolute deviation (average of |observed – predicted|) is 6.18 $\mu$g at an expected mass of 180 $\mu$g; the largest percent deviation (of average deviation) is 4.79% at 40 $\mu$g.

The validation of SOP 123 was successful; all acceptance criteria were met. Although there is some evidence of lack-of-fit to the linear model, the deviation from the predicted values is not large for the method's stated range of 50 to 150 $\mu$g. Thus, the final report for the validation of SOP 123 for use in

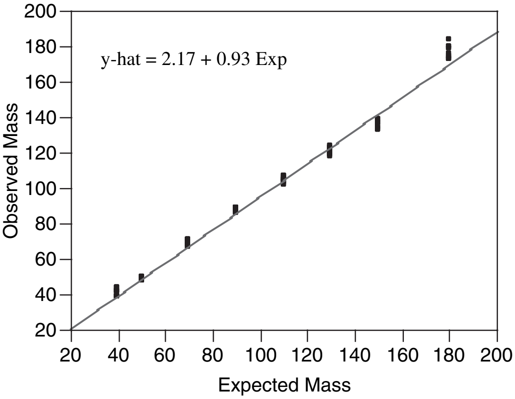


**Figure 7** Scatter plot with all the data with least squares line.

**Table 14** Lack-of-Fit Summary for Linearity of Example Validation Data with All Assays in the Linear Regression Analysis

| Lack-of-fit $P$-value | Maximum deviation (%) | Level of maximum deviation ($\mu$g) |
|---|---|---|
| <0.0001 | 10.31 | 40 |

measuring mass of Product W can state that the method is valid for its intended use.

## CONCLUSIONS

The quantitative method validation characteristics of accuracy, precision, and linearity were discussed in depth in this chapter. An example validation protocol is shown with a fractional factorial design matrix for the validation experiments. With careful planning, these three validation characteristics can be assessed from the same set of validation experiments. Thus, implementation of design of experiments can mean efficient use of resources. The analysis report was based on hypothetical data. Additionally, the author is neither recommending nor providing guidelines for what is deemed "acceptable" for the performance characteristics being assessed. The validation acceptance criteria should be based on sound scientific and business decisions as to what is required of the assay being validated.

One of the most important considerations in designing the experiments is balance over the factors of the SOP being validated. In the example, the validation assays were balanced over the factors operator, day, and expected mass. Lack of balance in a design can create erroneous statistical results and lead to inaccurate assessments of these validation characteristics.

## REFERENCES

1. Fry EM. The FDA's viewpoint. Drug Cosmetic Ind 1985; 137:46–51.
2. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human

Use. ICH Hamonised Tripartite Guideline Q2A: Text on Validation of Analytical Procedures. Geneva: ICH, 1994.

3. Torbeck LD. Assay Validation Basics. 2nd ed. Evanston, IL: Suffield Press, 2000.
4. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH Hamonised Tripartite Guideline Q2B: Validation of Analytical Procedures: Methodology. Geneva: ICH, 1996.
5. Pharmacopeial Convention. Chapter 1225: Validation of Compendial Methods in United States Pharmacopeia and National Formulary (USP 28-NF 23). Rockville, MD: United States Pharma Convention, 2004.
6. John, PWM. Statistical Design and Analysis of Experiments. New York: Macmillan, 1971.
7. Box GEP, Hunter WG, Hunter JS. Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building. New York: John Wiley & Sons, 1978.
8. Kuehl RO. Design of Experiments: Statistical Principles of Research Design and Analysis. 2nd ed. Pacific Grove, CA: Duxbury Thomson Learning, 2000.
9. National Committee for Clinical Laboratory Standards. Evaluation of the linearity of quantitative analytical methods; Proposed Guideline. NCCLS publication EP6-P. Villanova, PA: NCCLS, 1986.
10. SAS Institute Inc. SAS/STAT® User's Guide. Version 8. Cary, NC: SAS Institute Inc., 1999.
11. Searle SR, Casella G, McCulloch CE. Variance Components. New York: John Wiley & Sons, 1992.

2

# Validation of Chemical Reaction Processes

**ROBERT F. DILLARD**

Biostatistics and Data Management
Takeda Pharmaceuticals North America, Inc.
Lincolnshire, Illinois, U.S.A.

**DANIEL R. PILIPAUSKAS**

Global Manufacturing Services
Pfizer, Inc.
New York, New York, U.S.A.

## INTRODUCTION

A chemical reaction process, the large-scale conversion of raw materials to an intermediate or product, is considered validated if it can be shown to consistently produce material of defined quality following a set of manufacturing instructions. Attaining this knowledge and performance is highly dependent on acquiring a quantitative understanding of all critical factors that influence the process. The Food and Drug Administration (FDA) is developing ways to positively reinforce this concept. In the future, companies will benefit from a lower regulatory burden for process validation and improvement if their commercial manufacturing process can be shown to be well understood. In particular, a change control procedure would be sufficient to manage process improvements without submission of additional data to the FDA (1).

Chemical process development is about achieving this level of understanding under aggressive timelines and business constraints. Process developers typically do not have the luxury of knowing when process development must be completed and to what site it will be transferred. It is not uncommon for early toxicological and clinical results to lead to the decision to suspend development, only to be restarted months later when more promising results are acquired. When development work is suspended, the commercialization milestone is usually not delayed, which places tremendous pressure on the development organization.

Delivering the process under these constraints requires a balance between speed, information, and cost. Solutions lay on a continuum from performing process development the usual resource-intensive way, to creating new ways to acquire process knowledge that use a smaller number of resources. For timing and budget reasons, simply increasing the number of process developers is not always an option. Implementing highly parallel experimentation with robotic workstations may help, but can be impractical because of the need for equipment, specialized skills, and time to make this approach practical. And while one could meet the process development timelines by cutting corners, the company as a whole loses because of the less than optimal manufacturing process and, likely, higher cost of goods. The issue that the process developer faces, and this chapter addresses, is how to increase the amount of process understanding per experiment while minimizing the impact on the project timeline.

Achieving understanding requires the establishment of mathematical linkages between a complex set of inputs and process conditions (factors) with outputs (yield, impurity levels, etc.). Through a series of experiments where the inputs are varied systematically, relationships or models are established between the inputs and outputs. These relationships can range from purely empirical at one extreme to a fundamental or mechanistic understanding at the other. In the empirical approach the information gained tells us how the process works, but not why it works (e.g., yield goes up as factor x increases). Understanding why a process works and why it fails requires additional information. The investment, however, is returned in a more efficient learning process. Fundamentally, we propose increasing information through the integration of multifactor experimental designs with techniques for characterizing reaction kinetics. While the concept is described for the development of a chemical reaction process, it can be applied to the development of downstream unit operations as well.

## MULTIFACTOR EXPERIMENTAL DESIGN IN PROCESS DEVELOPMENT

Development of a chemical manufacturing process progresses through stages of increasing understanding of the underlying chemical and physical phenomena. These stages can be grouped into three broad categories: early development, characterization, and optimization. The characterization stage can be further subdivided into factor screening and range finding. These stages roughly follow the natural stages of learning: what factors are important, how do the factors relate to each other, and what is the optimum combination. At each stage, the experimental approach changes to reflect the type and depth of information being gathered.

Early process development starts with feasibility exploration. The purpose of this early stage is to confirm, at least at some high level, the reaction kinetics—what kinds of yield, purity, and reaction rates are possible. The predominant experimental design is the one factor at a time (OFAT) approach. When the synthesis scheme and reagents (i.e., recipe) become more clearly defined, development activities switch to a characterization phase. Here, multifactor experimental designs become the approach of choice to provide a quantitative link between process performance and process parameter levels. When this more complete level of process knowledge is achieved, experimental design becomes directed toward optimization, defining operational limits and raw material specifications. Table 1 outlines each of the key stages, their chief goal, and the predominant design approaches. In actual practice, these stages do not occur in a strictly linear fashion: they overlap. At times acquired knowledge will lead to process modifications and repetition of early stages.

**Table 1** Outline of Process Development

| Process stage | Goal | Design |
|---|---|---|
| Early dev. | Confirm synthesis route | One factor at a time |
| Characterization | | |
| Screening | Rank order driving factors<br>Determine rough reaction kinetics and by-products | Fractionated design<br>Plackett-Burman<br>Resolution III |
| Range finding | Refine process definition<br>Explore operating limits | Factorial designs<br>Resolution IV or better |
| Optimization | Optimize conditions for yield, time, impurity levels<br>Establish operating limits | Response surface<br>Central composite<br>Box Behnken |

This experimental design approach can be thought of as a lens that increasingly focuses the experimental effort in the optimal operating region. The process starts with limited, often qualitative knowledge about the reaction. The experimental designs sample the reaction space in a broad encompassing pattern looking for big effects. As the development process proceeds, the number of critical factors and their operating ranges become more clearly defined. The designs center around the reducing factor space with the spatial coverage becoming both more focused and more detailed. As the focus continues to narrow, the effort shifts to optimization, the spatial coverage fills in, and the models in turn become more complex.

## Design and Feasibility—Early Process Development

Development usually starts with the discovery synthesis, a process designed to produce only small quantities, and often involving chemicals and procedures not amenable to a manufacturing process. Early development converts the discovery route to a synthetic route that does not have chemical, safety, environmental, or operational issues that would prevent it from being commercially viable. This must be done before the drug substance solid form and impurity profile become set by formulation development and toxicological studies.

At this early development stage, there is often very little time available to study why the chemical reaction process works or what makes it fail. The process developer confirms the operability and adjusts the laboratory process as necessary. For example, are the reaction times and volumes necessary, are there better solvents, are reagent ratios appropriate, what are the gas and heat evolution rates, is mixing or solubility going to be a problem, can reasonable yield and impurity levels be achieved? At this stage, information about scalability may rest solely on reproducibility at a certain laboratory scale. Fortunately, scale up in the pilot plant under careful control and under the watchful eyes of the process developer usually results in successful production of larger quantities of material sufficient for clinical and toxicological studies, and dosage form development.

Experimental design during this phase of development is heavily weighted toward OFAT experiments. In OFAT experiments, only one factor is varied while all the other possible factors are held at some fixed condition. Pass–fail experiments dominate. Because these experiments are seldom replicated, only large responses to process factor changes are likely to be observed and causes for variation incompletely understood. There are, however, several key pieces of information that should be established at this stage before proceeding to characterization. Of primary importance is the establishment of the final synthetic route; key steps must be laid out and nominal working conditions established. There will be much detail to fill in, but the primary steps should be fixed.

## Characterization and Optimization—Commercial Process Development

Characterization starts once the synthetic route has been selected, although there will be opportunities to modify the route if the changes do not impact the final solid state or impurity profile of the final active pharmaceutical ingredient. The primary objective is to understand, through experimentation, the chemical and physical chemical processes involved in the transformation of raw materials to intermediates and products. The primary outcome is a process definition that includes the order of manufacturing steps, process parameter control methodology, process parameter limits, raw material specification, and diagnostic metrics.

It is at this stage of the development process that the use of design of experiments (DOE) brings much to the table. First, DOE is a very efficient means of covering the factor space—it maximizes the information from each set of experiments. By getting the most from each experiment the DOE approach allows the process developer to quickly focus on the critical process factors. In addition, by covering the factor space in a systematic way, DOE facilitates model building. Good designs ensure that each factor has sufficient levels to estimate model parameters and to test the appropriateness of the hypothesized models. DOE also brings a common structure to the experimental effort as it moves through each development stage. Comparisons can be made back to earlier experiments because history is maintained.

## GETTING TO A DESIGN

We will outline the key design and analysis features for each process development stage. Many of the detailed decisions inherent to creating a design will not be discussed here. Box, Hunter, Hunter (2) is a particularly good source for this information. All of the designs used here can be found in their text book. However, some general principles that apply to all stages are discussed first.

### Design Process

Any experimental design requires planning to be successful. Much has been written about this issue; see for example, Coleman and Montgomery (3). Table 2 captures the key features. The details of this planning are, for the most part,

**Table 2** Overview of the Design Process

1. Establish the objectives
2. Define any limitations
    Budget
    Lead times
    Physical resources
3. Identify the response variables and their relationship to each other
4. Brainstorm the possible factors and prioritize
5. Establish ranges for the factors likely to be in the design
6. Identify factors that will/will not be controlled. List settings of the controlled factors. For noncontrol factors can one block or randomize to lessen the effects?
7. Propose/select design
    What is the stage of development? What interactions are anticipated?
    Which design points are required to estimate responses (e.g., reaction rates)?
8. Establish roles/responsibilities and timelines
9. Execute

beyond the scope of this chapter, but two key steps, factor selection and response selection, require discussion here.

The most important step in the design process is step 4, the choice of factors. Their choice is more critical to success than the actual experimental design. A key input to their choice is the use of knowledge about the problem from a first principle's perspective. We start from the premise that chemical reactions follow certain fundamental laws associated with reaction kinetics. This suggests, for example, that it is not reactant concentration that is important but rather the ratio of reactants to the starting materials. It also suggests that time is always a critical factor, as is temperature. The list of first principle factors can be overwhelming. A critical part of the design process will be to prioritize the factors in order to keep the effort focused and the designs manageable.

Another important step in the design process is the choice of response(s) (step 3). Here again we look toward first principles for guidance. The primary responses will be the concentration/time curves associated with the products' and by-products' formation as well as the starting material loss. From these curves, we will determine the yield and purity levels that can be achieved and, more importantly, the reaction rates. The time curves are the starting point for understanding the underlying reaction kinetics. By analyzing how the concentration of starting materials, products, and intermediates change over time, we will understand the reaction mechanism(s). Notice here a key feature of the design planning—there are multiple responses to be selected, not just sampling times, but different time profiles that measure different features of the reaction.

Fundamentally, the goal is to maximize the chemical yield while minimizing the formation of by-products. Fortunately, factor combinations that give higher yield also tend to give lower levels of the unwanted materials. However, there are trade-offs to be made among the by-products, with some needing to be more strictly controlled than others. The optimum factor levels as a result are compromises between competing needs. These competing needs mean that, often, more factors than just those effecting yield need to be carried forward.

## Increasing the Information Content

As we have noted, combining a list of factors into a design first requires consideration of the development/validation stage. As the goals of each stage change, the design choice changes to accommodate the goals. We will illustrate this in some detail but a common feature, regardless of the development stage, is the heavy reliance on factorial designs. In these designs, each factor is tested at two or more equally spaced levels. The experimental design considers the factor levels at all possible combinations. As the number of factors and their levels increases, the number of possible combinations can get overwhelming. But through the use of clever fractionation—that is, leaving out some factor combinations—one can make the experimental effort manageable and still insure that key effects can be estimated.

However, process chemistry has several features that distinguish it from other DOE applications. As noted, time is always a critical factor. Time can also be a response, as in time to reaction completion, but its primary role is as a factor with carefully chosen levels. Generally, time is considered a *nested factor*; the levels selected depend on the level of other factors in the experimental run. Reactions proceed at very different rates depending on the levels of the other factors, especially temperature, thus the time points providing the most information will differ from experiment to experiment. One can think of time as a factor that is *subsampled* from within the other factor combinations in the design. In fact, the design is usually built in a sequential fashion. First, the nontime factors (whole plot factors) are combined into a design, then the levels of time are determined separately for each experimental run (whole plot combination). In order to facilitate comparisons, one may choose to select some time points that are common across the experiments, but the key criteria is to choose time points that build toward quality estimates of the main responses—the reaction rates.

Once a design has been established, one has to consider in what order to perform the experimental runs. The natural tendency is to reduce costs by running experiments in groups that are somewhat alike in the factor settings. For example, if an experiment included temperature, it would be natural to want to run all of the high-temperature experiments at one time. Avoid this kind of short cut. A fundamental tenet of experimental design is that the experimental runs are performed in a random order. This will ensure that the factor effects are not biased by time or environmental trends. More importantly, randomization ensures that replicate experiments represent the true variation in the system. Use of a formal randomization plan that makes use of a random number generator is encouraged.

The time factor, however, is typically not randomized. Once a reaction is set up and going, it is repeatedly sampled in time—one does not typically start over with a new reaction for each new time point. Time then is referred to as a *repeated factor*. (With the increasing availability of on-line monitoring technology, continuous sampling has become quite practical, but there still are reactions that cannot be repeatedly sampled. Though even in these cases one often relies on surrogate measures that can be repeatedly sampled, e.g., hydrogen uptake.) This restriction on the randomization leads to a subclass of designs referred to as split-plot designs. In split-plot designs, the *whole plot* factor levels are randomly selected and then the whole plot is sampled or split and additional *split-plot* treatment factors applied. The terminology arises from the original development of these designs in agricultural experiments. The whole plot here refers to each combination of nontime factors that make up each experimental run. The split-plot factor, time, is applied to each whole plot.

Because of the lack of randomization in time, the variation in the data observed from time point to time point within a reaction (whole plot) is likely to be less than that observed from reaction to reaction. In effect, two process variance terms have been introduced, one reflecting the within-reaction variation, and a second reflecting the variation from reaction to reaction. This has consequences for how to determine which factor is having effects beyond the inherent process variation. Fundamentally, any factor effect measures need to be compared against the correct comprehensive variance estimate.

Another key feature of the application of DOE to process chemistry is that many of the critical factors, especially time and temperature, are nonlinear in nature. They enter the model in a nonlinear way (Appendix 1). This has important consequences for the spacing of factor levels (step 5). A flexible approach is needed: one that allows for factor levels that will provide good information on reaction rates and provide for a broad sampling of the response region. In some situations, it may be possible to describe an optimum factor spacing, but this optimum will depend on knowledge of the parameters themselves—information we seldom have. Rather than focus on the optimum, a few useful principles can be applied to choose reasonable levels. First, when selecting levels for the split-plot factor time, do not be stingy. Multiple sample times are generally cheap compared with the cost of the experiment. More is better as long as one does not sample so frequently as to change the nature of the reaction. When spacing the samples, try to get coverage in the following three key regions: (*i*) the initial reaction phase, (*ii*) the reaction completion phase, and (*iii*) the postreaction or hold time phase. Coverage in these regions will allow for sound estimates of the rates of the reaction for both the product and by-products. For the whole plot factors, do not be timid in the choice of factor ranges—especially early in the experimental plan. Strive for poor results as well as good. Do not try to focus on an optimum region too quickly. Instead strive to understand what can go wrong as well as right.

## Analysis Methods

The analysis methods proposed here take advantage of kinetic modeling approaches as well as more traditional linear models, such as analysis of variance (ANOVA). At the screening stage, because so few time samples are taken, one typically condenses the concentration/time curves from each run into summary (whole plot) measures, such as yield and by-product formation. Simple analysis approaches, such as normal probability plots or ANOVA, can then be applied. Visual inspection of the concentration–time curves is still recommended to identify clues to the reaction mechanism. If sufficient sampling times are available, attempts to fit a rate law to the data may be profitable.

As the DOE stages progress, the time dimension is folded into the models. Initially, this can be done through the use of linear (mixed) models, including response surface models. But early on we push for kinetic models as a hypothesis for the reaction. These models are often simplified to allow for rate estimation under sparse design scenarios and are usually fit to each measured species in each experiment (reaction) separately. One obtains reaction rates for each species and then applies usual ANOVA techniques to the estimated rates. The focus is on how these rates change as the key factors change. Are the rates temperature dependent, are they dependent on the purity of the starting material, are they reproducible? Reproducibility is a key goal—we want to get an early estimate of the experimental variation, particularly the between-experiment (whole plot) variance. Later, as the knowledge expands, kinetic models are fit to all reactions simultaneously. The proposed models are examined for how well they fit the data. The kinetics are either confirmed or new models hypothesized.

Box and Youle (4) pointed out the mathematical links between the more common empirical linear models used in many DOE applications and the fundamental reaction kinetic models proposed here. However, the kinetic model approach has several advantages. First, it communicates the results in a more chemically intuitive and informative way to process chemists and engineers. Second, it improves the quality of the model fit—it better deals with the underlying nonlinearities—and, more importantly, better predicts outside of the observed range (projections of yields greater than 100% are not very useful). Another very important advantage is that the proposed rate expression, if correct, can predict what the experimental results should look like—a type of reference point that prevents one from just accepting the results from an experiment. If the proposed rate expressions and new data do not agree, it is a clear indication that something is amiss—it is a very effective way to become a skeptic about the exact amount of understanding one has. A mechanistic model forces the experimenter to really challenge their logic and test their level of understanding. You cannot do that with standard linear models alone. That is not to say that linear models do not have their place—they do. We will illustrate the use of ANOVAs, for example, in our application, but our goal is to establish mechanistic-based models.

## STAGES OF PROCESS DEVELOPMENT

We have been discussing aspects of individual designs but, as noted, process chemistry development is a series of experimental designs. One should budget and plan from the beginning with this in mind. We will now move to describe each of the key stages in more detail.

### Factor Screening

The basic approach in factor screening (Table 3) is to stress or perturb the process over a broad range of the factor space to understand how it responds with respect to yield and impurity levels. Conditions that generate poor yields and high impurity levels are actually sought in this stage.

Developing a list of factors to screen can be a daunting task, given the potentially large number of factors and the challenge to test the minimum number without missing an

**Table 3** Screening Stage Summary

| Stage | Screening |
|---|---|
| Trigger | Synthesis identified, need to identify important factors |
| Goal | Rank process factors (main effects) |
| Issues | Optimizing before factors established/failure modes identified. Factor ranges potentially too small, the effect too small. Still tinkering with process |
| Approach | Follow effects hierarchy. Study one unit operation to minimize experimental error. Do not isolate product—measure species in process stream |
| Design | Highly fractionated. Pilot a block to ensure good factor range selection |

important factor. But the selection process need not be approached with trepidation if the factor identification is made within a physicochemical framework. By basing factor selection on principles of mechanistic organic chemistry, physical organic chemistry, and physical chemistry, a prioritized list of the most influential factors can be developed. Given the need for in depth knowledge in multiple disciplines, the use of subject matter experts to assist in factor selection is recommended.

Obtaining reliable data will be highly dependent on factor range selection, that is, defining where in factor space the experiment will be conducted. Unfortunately, ranges are often approached timidly. There is a fear of an experiment failing and wasting time. This often results in factor ranges being too narrow to provide a response larger than the experimental error. Be bold—there will be opportunities to narrow ranges later. Factor ranges should be broadened to the point that there is a reasonable degree of certainty that the experiment will still provide meaningful results. Also, understanding what physical laws may be controlling the process helps in range selection. For example, for chemical reactions, temperature influences the rate exponentially, whereas concentration typically influences the rate linearly.

The main goal at this stage is to rank order the factors so that future studies can focus on the key ones. The design tools of choice are the class of highly fractionated (usually saturated) designs, especially resolution III designs. These designs are sometimes called main effects designs, because main effects are all they are really capable of estimating. Nevertheless, this class of designs excels at screening a lot of factors with minimal investment in time and resources. A principle feature of resolution III designs is that the number of experiments is only one more than the number of factors—you do not get cheaper than that. Table 4 illustrates a resolution III design in which seven factors are studied in eight experimental runs.

To build a resolution III design, first write out the effect matrix associated with the full factorial design whose size (i.e., number of runs) is just greater than the number of factors to be screened. Then assign each screening factor to a

**Table 4** Effect Matrix for $2^3$ Design

| | Factor levels | | | Factor levels based on interaction effects | | | |
|---|---|---|---|---|---|---|---|
| Run | 1 | 2 | 3 | 1*2 (→4) | 1*3 (→5) | 2*3 (→6) | 1*2*3 (→7) |
| 1 | – | – | – | + | + | + | – |
| 2 | + | – | – | – | – | + | + |
| 3 | – | + | – | – | + | – | + |
| 4 | + | + | – | + | – | – | – |
| 5 | – | – | + | + | – | – | + |
| 6 | + | – | + | – | + | – | – |
| 7 | – | + | + | – | – | + | – |
| 8 | + | + | + | + | + | + | + |

*Note*: +, high level of the factor; –, low level of the factor.

column in the effects matrix, including the interaction terms. Consider the example in Table 4 where we have seven factors to screen. The full factorial design whose size just exceeds 7 is $2^3$. The signs for the interaction term effects normally associated with this design have been used to code for the additional factors 4, 5, 6, and 7.

If there are fewer factors than columns then some key interaction conditions can also be estimated. Simply do not assign any treatment to the interaction(s) of interest. For example, if in Table 4 there were only six factors in the design, then one of the columns could be used to focus on an interaction term of particular interest. Suppose an interaction between factors 1 and 2 was suspected, then one could leave the column associated with factor 4 unassigned, as it is confounded with the interaction between factors 1 and 2. Instead, factors 4 to 6 would be assigned to columns 5, 6, and 7. One could use this same technique to get a crude estimate of the underlying experimental variation by associating the effects of unassigned columns with noise. Using unassigned columns to estimate either interactions or variation should be done with a great deal of caution, as there is much confounding in resolution III designs.

The number of runs ($n$) associated with resolution III designs come in $2^k$ increments (4, 8, 16, and so on). Each can handle up to $n$–1 factors. The choices of designs can be expanded to handle intermediate numbers of runs through the use of Plackett-Burman designs. These saturated designs handle run sizes that are multiples of 4, namely $n = 12, 16, 20, 24$, and so on. This allows for additional design choices. The $n = 12$ and $n = 20$ designs are particularly useful. Again unused conditions can be used to include key interaction terms or kept as dummy placeholders to be used to obtain an estimate of variation.

In keeping with the efficiency of the screening design we tend to choose a relatively small number of (split-plot) samplingtimes for each (whole plot) design point (keeping in mind that sampling multiple times within a reaction is relatively cheap). A rough rule of thumb is to include at least one sampling time point in each of the three key reaction regions described earlier. The first region corresponded to the initial rate of the reaction. If possible take a sample approximately where the reaction is half way to its anticipated maximum yield, although, depending on the speed of the reaction this can be difficult to do. The second key region corresponded to the reaction completion—the point of maximum yield. The third sample region was the postcompletion region. Here, try to sample well after the reaction is complete when the by-products begin to dominate. If additional time points are possible, the highest priority locations will include additional samples near the reaction completion point, followed by additional samples in the initial phase of the reaction.

The statistical analysis approaches associated with screening designs rely on a certain lack of statistical rigidity. Fundamentally, the goal of the analysis is to start to develop an understanding of the temporal changes for the raw materials, reaction intermediates, and product. One is looking for which factors have the most impact on these time profiles. It can be difficult at this early stage to formally include time as a variable in a statistical analysis. As we have noted, time is a nested factor, thus the sample times for each reaction are often unique. This makes it difficult, for example, to code for time in an ANOVA analysis. We recommend collapsing the time dimension into what might be called a whole plot response and then performing a first pass analysis using the whole plot factors. There are a couple of ways to collapse. If some time points are common to each experiment, then one can analyze the whole plot effects separately at each time point—essentially slicing by time. This can be especially helpful when one is interested in the whole plot factor effects on the initial reaction rates. Another approach is to summarize or average the response(s) over the time dimension into one response. So, for example, one could analyze the maximum observed concentration across time (i.e., maximum yield). For by-products, one might perform an analysis using the last observed concentration from each reaction. Finally, one can also treat time as a response, as in time to maximum yield.

A key limitation in the analysis of screening designs is the lack of replication. Thus, an estimate for the underling variation is typically not available, and there are no clear factor candidates to combine into an error term. Nor is there likely to be historical information at this stage from which to approximate a variation estimate. This makes ANOVA methods difficult to implement. But there are statistical methods that can help. One of these is normal probability plots. These plots graphically determine which factors are influential by rank ordering the factor effects (a useful exercise in itself) and then plotting the effects on normal probability paper. Effects that are significant, not just process noise, will stand out as not following an overall linear trend. However, the label of "significant" should be used cautiously if at all. At best, the underlying variation is poorly estimated and the ability to distinguish factor effects from those effects driven by interactions will be poor.

### Example

Appendix 1 lays out a simple hypothetical chemical reaction, which we will use to simulate a series of development experiments and the approaches to their analysis and information accumulation. In the hypothetical example two reactants, A and B, combine to form product C. Two undesirable by-products, D and E, are also formed. The goal is to maximize yield of C while minimizing the by-product levels. In this example, we will strive to find conditions that keep D and E under 1.0%.

Imagine that we are just starting the screening stage of process development. We have received the chemical process from discovery and completed our early process development work. OFAT experiments have established the fundamental chemistry to be sound and reasonably reproducible. We have established a short list of the solvents and reagents, crude ratios, volumes and times, and have identified key impurities. We are ready to develop the first screening DOE.

The team has brainstormed the factors and developed a prioritized list of the key ones. For each factor, ranges were established that were well outside of the anticipated operating range in order to maximize the effect size (but not so large as to change the fundamental reaction chemistry or overly stress the process facilities). Table 5 summarizes the final list of factors and their ranges that are to be included in the first DOE. Time, of course, will also be a (nested) factor, but, as previously discussed, its levels will be determined after the whole plot design is established.

An eight run, resolution III design, was developed for the whole plot factors using the approach described earlier (Table 4). Sampling times were then brainstormed. For this first formal experiment, there was not a lot of information on

**Table 5** Key Factors and Ranges—Screening Stage

| | Ranges | |
|---|---|---|
| Factors | Low (–) | High (+) |
| X1: Ratio of A to B | 1:1 | 1:2 |
| X2: Temperature | 25° | 50° |
| X3: Addition rate, B | 1000 u/hr[a] | 1 u/hr |
| X4: Ph | 7.5 | 8.5 |
| X5: Solvent type | S1 | S2 |
| X6: Agitation level | Low | High |
| X7: Reactant, B, source | B1 | B2 |

[a] 1000 units/hr represents a batch mode of operation.

the sensitivity of the reaction rates to temperature, so it was decided to use the same sampling times for all reactions. The times selected were half, one, two, three, and eight hours. The reaction was expected to finish in the two- to four-hour time window, so both times were included in the hope that at least one would be near the maximum yield associated with each experiment. Two sampling times were included early in the reaction (half and one hour) in order to get some information on the reaction rate(s). The eight-hour sample time was included to establish whether a hold period at reaction completion was appropriate and to force degradent formation. Table 6 summarizes the whole plot design (factor levels are given in coded units), and several key whole plot summary responses. The design and results are given in standard order—the experiments were actually performed in random order.

Figure 1 contains the actual observed time profiles from each of the eight runs. Much can be learned just by examining the graphs. Foremost is that the chemistry is reasonably well behaved. Yields are good, and except at high temperature (bottom half of the figure) the by-product levels appear to be close to our goal (<1.0%). Both the rapid addition (left side of the page) and the slower method of addition (right

**Table 6** Resolution III Design of Experiments—Factor Screening Experiment

| | Factors | | | | | | | Key responses | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Run | X1 | X2 | X3 | X4 | X5 | X6 | X7 | $C_{max}$ (%) | Rate (u/hrs) | D@8 hr (%) | E@8 hr (%) |
| 1 | – | – | – | + | + | + | – | 87.1 | 0.11 | 3.6 | 0.4 |
| 2 | + | – | – | – | – | + | + | 96.7 | 0.24 | 4.1 | 5.4 |
| 3 | – | + | – | – | + | – | + | 88.9 | 0.22 | 24.6 | 0.2 |
| 4 | + | + | – | + | – | – | – | 92.8 | 0.46 | 26.0 | 35.5 |
| 5 | – | – | + | + | – | – | + | 86.0 | 0.11 | 3.8 | 0.5 |
| 6 | + | – | + | – | + | – | – | 102.0 | 0.25 | 4.6 | 6.8 |
| 7 | – | + | + | – | – | + | – | 95.1 | 0.48 | 24.5 | 0.4 |
| 8 | + | + | + | + | + | + | + | 101.4 | 2.03 | 25.8 | 38.9 |

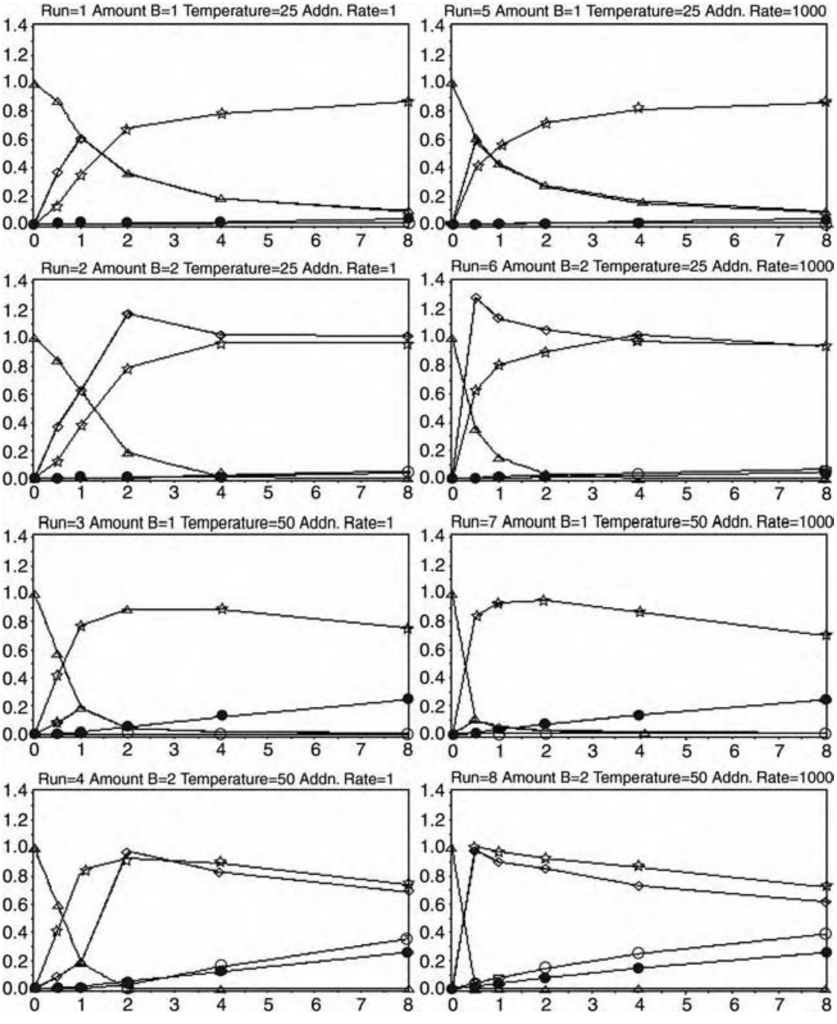


**Figure 1** Time profiles from resolution III design of experiments. *Note*: Vertical axis is concentration, horizontal axis is time; concentration $A = \Delta$; concentration $B = \Diamond$; concentration $C = ☆$; concentration $D = ●$; concentration $E = ○$.

side) provide similar yields, although the method of addition is of course a slower reaction. Some striking features exist in runs three, seven, four, and eight that are worth further discussion.

By simple inspection of the plots for runs three and seven, the rising level of the by-product D appears to be accompanied by a decreasing level of the product. This correlation can be an important clue to the formation chemistry of D and is certainly an important observation concerning the stability of the product C. Inspection of plots for runs four and eight reveals another correlation. Formation of the by-product E seems to increase with a corresponding loss of starting material B, possibly providing a useful clue to formation of E and the stability of B in the reaction mixture. Insights like this into the reaction mechanism illustrate the importance of making multiple observations during each run.

We now bring more formal analysis approaches to bear on this data. Our primary goal is to rank order our factors. We actually have enough information to begin kinetic modeling, but we will start with a more empirical approach to the analysis and develop the kinetic modeling approaches more fully in the next factor range finding experiments. Four primary responses (Table 6) were used to describe the time profiles from each run: (*i*) yield ($C_{\mathrm{max}}$), as measured by the maximum observed concentration of C expressed as percent of the starting concentration of A, (*ii*) the reaction rate for the formation of C (rate), measured in units per hour as the maximum observed concentration of C divided by its corresponding time, (*iii*) the level of by-product D at the eight-hour time point (D@8 hr), and (*iv*) the level of by-product E at the eight-hour time point (E@8 hr), both expressed as a percent of the starting concentration of A.

The analysis utilized the approach of normal probability plots, as no reliable historical estimate of variation was available. For each primary response the observed factor effects were plotted on normal probability paper in rank order (Fig. 2). Here, "factor effects" are estimated as the difference of the average response at the high level (+) from the average response of the low level (–). On these plots, values that deviate markedly from the general trend line indicate significant effects. The largest factor effects (in absolute value) are labeled.

Note from the figure there are two clear dominant factors, the ratio of A to B (factor X1) and temperature (factor X2),

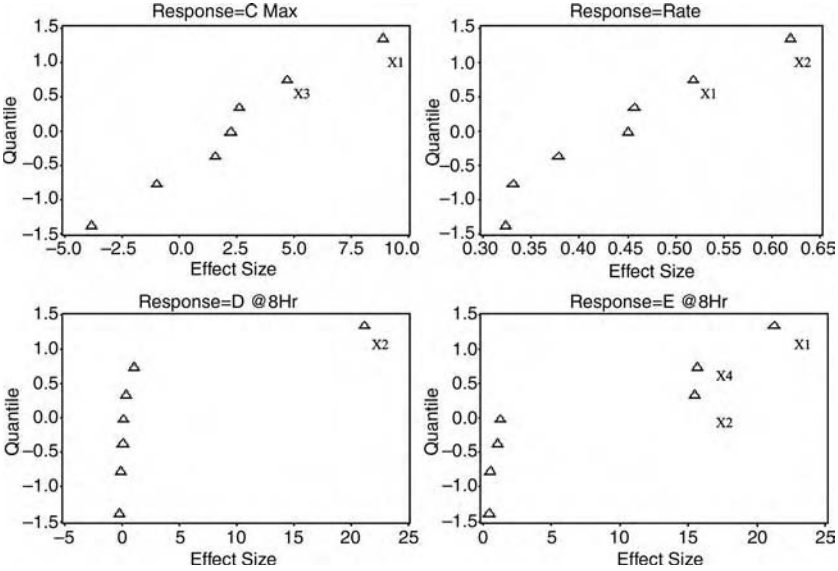


**Figure 2** List of effect estimates w/normal prob. plot.

each having different importance depending on the response. There are a few factors in the middle of somewhat unknown importance, and several having very little impact. Ph (factor X4), appears to be important to the formation of E, but note this factor is confounded with the interaction between factors X1 and X2—the interaction between the ratio of A to B and temperature. Temperature does not have much impact on yield ($C_{max}$) although it does affect the reaction rate and has a substantial impact on by-product formation. The challenge of the next set of experiments will be to better quantify the effects of the important factors and unconfound them.

## Range Finding

As we move out of screening into range finding, the goal shifts. The sparse designs used in factor screening helped us to identify the key factors to carry forward. They allowed for a rough estimate of the range of reaction rates and a better understanding of the level of by-product formation. Based on the results, the factors can now be ranked by strength of experimental response and the important ones selected for further exploration. In range finding the experimental effort shifts toward more focus on a smaller number of key factors, with emphasis on two-factor interactions and improving the accuracy and precision of the parameter estimates (Table 7). The goal in this work is to further test our understanding of how the chemical reaction process works, support operating limit selection for important operating factors, control limits, and registered operating limits for the critical process factors and specification for raw materials and product. This is where new conditions or combination of conditions will be tried to get at potential interactions. How thoroughly we study this stage determines how many future plant issues are avoided before manufacturing. Process knowledge gained here will set the stage for the optimization efforts to come.

The major design type used are two-level factorials, either in full form ($2^k$) or fractional form ($2^{k-p}$). These designs have several advantages over other approaches. First, like the screening designs, they are very efficient: the variances associated with the parameter estimates are as small or nearly as small as those from any other design of the same size. They

**Table 7** Range-Finding Stage Summary

| Stage | Range finding |
|---|---|
| Trigger | Factors ranked. Is performance predictable/does it follow chemist model/are the assumptions correct? |
| Goal | Identify two-factor interactions. Establish more accurate mathematical linkage between performance and factor settings. Set groundwork for limits and spec setting and optimization. Identify critical process factors |
| Issues | Inferences/assumptions about untested factor settings. Is the laboratory scale a good surrogate for manufacturing scale? Range selection: make it work vs. running into a failure vs. nonrobust response (missed information) |
| Approach | Study one unit operation to minimize experimental error. Do not isolate product measure species in process stream |
| Design | Fractional factorial with some replication |

also allow for simple straightforward calculations of the effects. The designs form a hypercube, a relatively simple shape that facilitates understanding and presentation of the results. The designs can also be easily expanded to include additional fractions or broader factor ranges.

Several practical considerations come into play at this stage of development. Foremost is the practical limit on how many experiments a lab can reasonably manage in one study. Given the complexity of the staging and the reaction times required for even simple process chemistry, our experience suggests that one is unlikely to be able to manage more than 16 to 20 experiments at a time. Good design choices that match these kinds of numbers are the full $2^4$ design, and the fractionated $2^{5-1}$, and $2^{6-2}$ designs. The $2^{5-1}$ design is particularly well suited for range-finding needs. It has the nice property that main effects are confounded only with four-factor interactions and two-factor interactions are confounded only with three-factor interactions. If we assume that higher order interactions are negligible we can estimate the five main effects and all 10 two-factor interactions with only 16 runs. Augmenting these designs with multiple repeats of a center point is highly encouraged. The center points provide valuable information about curvature in the responses as well as provide the first reliable estimate of process variation.

Often, either because there are still a lot of factors to screen, or because the number of experiments is costly, or just because too much is still unknown, the work will proceed cautiously, in stages, using blocks or “foldover” designs. These approaches first implement highly fractionated, typically resolution III designs, and then, depending on the results, augment the initial designs with additional blocks of runs. So, for example, if the results are promising then the design can be augmented with additional runs to make it more complete. If the initial results are not good then the design can be shifted to a new region by changing the factor ranges or adding new untested factors to the design. In this way, energy is minimized in poor regions of the factor space.

Factor range selection will start from what was learned in the screening stage, and typically narrow the ranges. Avoid the temptation to narrow too much or try to focus too quickly on a perceived optimum area. Again, subject matter expertise is vital in making good choices. A critical consideration to range selection at this stage is some knowledge of the manufacturing operation range. As a rough rule of thumb, we strive to have the ranges be at least twice as broad as the anticipated operating range.

As with factor screening, the time dimension is applied after the whole plot design is established. The time dimension is more thoroughly sampled at this stage, often with two or more sampling times in each of the three key regions. Additional samples collected from the initial reaction phase allow for a better understanding of the reaction rates, in particular whether there are any lags in the rates related, for example, to the need for adequate mixing. Samples taken near the anticipated reaction completion provide information on the "flatness" of the yield profile. This will be valuable information as we look toward defining the reaction hold time. Finally, we often extend the reactions far past their completion points in order to provide information on the rate of by-product formation and in particular to gain information on degradation product formation.

The analysis approaches for range finding are more formalized than those used under screening. Here, we begin to develop reliable variance estimates allowing for more reliance on statistical significance to assist in the identification of key effects (although the power of the design must be considered). The use of the ANOVA technique is common, particularly as a first pass analysis to determine which factors and interactions are most influential. We again have the challenge of dealing with the time factor. As we did in factor screening, a good first pass technique is to collapse on the time dimension and then perform an ANOVA analysis on these whole plot responses.

The use of models that incorporate the time dimension develop at this stage. Models can be empirical (e.g., polynomial-response surface) or mechanistic in nature. We believe in getting to a mechanistic-based model as early as possible. Mechanistic models best integrate time and best communicate-the chemical understanding. However, mechanistic models can get very complex and difficult to fit. Simplified models that fit separately to each experimental reaction can help keep the analysis manageable. The parameters often have high standard errors early in the game when the time profiles and conditions are limited. But even the use of less than perfect models early can help to build insight.

### Example

Consider the results from our first screening experiment. We established, not surprisingly, that the ratio of A to B, the temperature of the reaction, and of course time were all critical factors. The third factor, addition rate, did not appear to be that influential, although it may have some effect on yield. We were less certain about the fourth factor, Ph. It had a marginal effect on the formation of the E by-product, but this effect might be better explained as an interaction between the reactant ratios and temperature. The team decided to carry all four of these factors forward, and brought one new one to the table—the use of a catalyst. The factors and their ranges are given in Table 8.

The (whole plot) factor ranges were adjusted from those used in the first DOE. For the ratio of A:B and for temperature, the maximum level was lowered in order to make the range of prediction more in line with the capabilities of the manufacturing site. For example, the 50° maximum used in the first screening DOE, although very useful in that context, is now too far from or anticipated operation temperature

**Table 8** Key Factors and Ranges—Range-Finding Stage

| | Ranges | | |
|---|---|---|---|
| Factors | Low (–) | Mid (0) | High (+) |
| X1: Ratio of A to B | 1:1.0 | 1:1.3 | 1:1.6 |
| X2: Temperature | 25.0 | 32.0 | 40°C |
| X3: Addition rate, B | 1000.0 | 2.0 | 1.0 u/hr |
| X4: Ph | 7.5 | 8.5 | 9.5 |
| X5: Catalyst | 0.0 | 0.5 | 1 |

of 25° (i.e., room temperature). The maximum temperature was reduced to 40°. The maximum concentration of B was reduced to a ratio of 1:1.6. These new ranges are still well outside of our anticipated operating range—but we want large, "modelable," effects. The addition rate did not seem to have a large effect, but it was felt that we should continue to collect data from both operation modes (batch—1000 units/hr and method of addition—1 unit/hr). These ranges were not appreciably changed. The mid-level of 2 unit/hr gives an addition time of approximately one half-hour. Because we were uncertain about the pH effects, its range was widened. The mid-value for each factor listed in Table 8 will be used to create a center point for the design.

The sampling times were extended to 12 hours, as some reactions in the initial experiment did not appear to be complete. In addition, sample time points at 10 and 20 minutes were added to the high temperature reactions to allow for better estimation of the reaction rates. Note the nesting principle applied here—higher temperature experiments require different sampling times.

A $2^{5-1}$ design with two center points was planned and executed. The design and summary results are given in Table 9. The time profiles are too numerous to be given here; but, as with the first DOE we can condense the information into primary whole plot responses. These are essentially the same responses as used in the screening experiment with the addition of two new ones—the level of by-products, D and E, observed at the time ($T_{\text{max}}$) of maximum yield. In other words, at or near the reaction completion point, what were the observed levels of the by-products?

A first pass analysis was conducted on the primary responses. Note that because we have replicated the center point we now have an estimate (although crude) of the underlying variation allowing for the use of the ANOVA approach to the analysis. An ANOVA model that included all main effects, two-factor interactions, and a curvature term was fit to each of the five key responses. The results are summarized in Table 10. Tabled is the estimated effect for each of the main effects along with a flag indicating the magnitude of the

**Table 9** 25–1 Design of Experiments—Range-Finding Experiment

| | Factors | | | | | Key responses | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Run | X1 | X2 | X3 | X4 | X5 | $C_{max}$ (%) | Rate (u/hr) | D-$T_{max}$ (%) | D-12 hr (%) | E-$T_{max}$ (%) | E-12 hr (%) |
| 1 | – | – | – | – | + | 90.7 | 0.08 | 5.9 | 5.9 | 0.4 | 0.4 |
| 2 | + | – | – | – | – | 96.1 | 0.08 | 6.7 | 6.7 | 3.5 | 3.5 |
| 3 | – | + | – | – | – | 92.7 | 0.23 | 5.6 | 18.5 | 0.2 | 0.3 |
| 4 | + | + | – | – | + | 100.8 | 0.50 | 2.6 | 19.4 | 1.0 | 10.9 |
| 5 | – | – | + | – | – | 90.5 | 0.11 | 3.8 | 6.0 | 0.5 | 0.5 |
| 6 | + | – | + | – | + | 97.5 | 0.12 | 4.6 | 6.5 | 2.9 | 4.0 |
| 7 | – | + | + | – | + | 94.4 | 0.47 | 3.0 | 19.3 | 0.4 | 0.5 |
| 8 | + | + | + | + | – | 97.7 | 0.49 | 3.5 | 20.1 | 2.8 | 12.2 |
| 9 | – | – | – | + | – | 89.7 | 0.07 | 6.0 | 6.0 | 0.4 | 0.4 |
| 10 | + | – | – | + | + | 98.8 | 0.12 | 4.0 | 6.5 | 2.4 | 3.3 |
| 11 | – | + | – | + | + | 90.2 | 0.30 | 3.9 | 18.6 | 0.3 | 0.2 |
| 12 | + | + | – | + | – | 98.7 | 0.33 | 4.3 | 19.5 | 2.1 | 10.6 |
| 13 | – | – | + | + | + | 90.9 | 0.08 | 5.6 | 5.6 | 0.5 | 0.5 |
| 14 | + | – | + | + | – | 98.5 | 0.12 | 4.5 | 7.0 | 3.0 | 4.0 |
| 15 | – | + | + | + | – | 91.1 | 0.30 | 4.9 | 19.3 | 0.4 | 0.5 |
| 16 | + | + | + | + | + | 97.7 | 0.98 | 1.7 | 21.4 | 2.0 | 11.5 |
| 17 | 0 | 0 | 0 | 0 | 0 | 96.1 | 0.32 | 2.3 | 11.0 | 0.9 | 2.1 |
| 18 | 0 | 0 | 0 | 0 | 0 | 94.3 | 0.23 | 3.5 | 12.0 | 1.1 | 1.9 |

**Table 10** Analysis of Variance Summary

| | Effect size | | | | | |
|---|---|---|---|---|---|---|
| Factor | $C_{max}$ (%) | Rate (u/hr) | D-$T_{max}$ (%) | D-12 hr (%) | E-$T_{max}$ (%) | E-12 hr (%) |
| X1 | 6.9[a] | 0.14 | –0.9 | 1.0 | 2.1[b] | 7.1[c] |
| X2 | 1.3 | 0.35[a] | –1.5 | 13.3[b] | –0.6[a] | 3.8[c] |
| X3 | 0.1 | 0.12 | –0.9 | 0.5 | 0.3 | 0.5[a] |
| X4 | –0.6 | 0.03 | –0.1 | 0.2 | –0.1 | –0.1 |
| X5 | 0.8 | 0.11 | –1.0 | 0.0 | –0.4 | –0.1 |

[a] $P < 0.1$.

[b] $P < 0.05$.

[c] $P < 0.01$.

$P$-value associated with a test for the significance of each factor. $P$-values less than 0.05 indicate a strong effect. Values less than 0.10 (or perhaps 0.15) indicate factors that we should continue to track and use in our model building. Note that the table does not list the interaction effects. This is because for nearly all the responses none of the interactions were significant (at $\alpha = 0.1$). Only the amount of E at 12 hours had significant interactions, with X1*X2, X1*X3, and curvature all having significant effects.

The ANOVA clearly establishes that the predominant effects are driven by the first three factors (ratio A:B, temperature, and addition rate) and the lack of effects associated with X4 or X5 (pH and catalyst). The A:B ratio (factor X1) effects yield as measured by $C_{\text{max}}$ with the yield increasing by about 7% as one moves from the low level of the A:B ratio (1:1) to the high level (1:2). Unfortunately, the amount of E formed at 12 hours also increases by about 7% as the level of the A:B ratio increases. Temperature (X2) effects the formation of both by-products and the rate of the reaction. The amount of D formed is particularly sensitive to temperature, increasing by an average of 13% as the temperature is raised from its low level of 25° to its high level of 40°. The addition rate (X3) effects the formation of E. Interestingly, the amount of D formed near the reaction completion point (D@$T_{\text{max}}$) is unaffected by any of the factors. Unfortunately, its levels are all too high to meet our goal of by-products less than 1% (Table 9), suggesting that it may be necessary to compromise yield.

To better understand the magnitude and shape of the factor effects it is helpful to examine effects plots. These plots depict a response ($y$-axis) for factor(s) of interest ($x$-axis) averaged over the remaining factors of the design. Figure 3 provides key effect plots for yield and the formation of D and E at the 12-hour hold point.

The effects plots clearly show the effects of the A:B ratio on yield, with higher levels of B leading to higher yields. But, as already noted, higher levels of B also lead to higher levels of the E by-product (lower left graph). Higher temperature leads to higher levels of by-products—both D and E (upper right and lower left graphs). However, the plots clearly illustrate the different nature of the effects. Recall that the E by-product

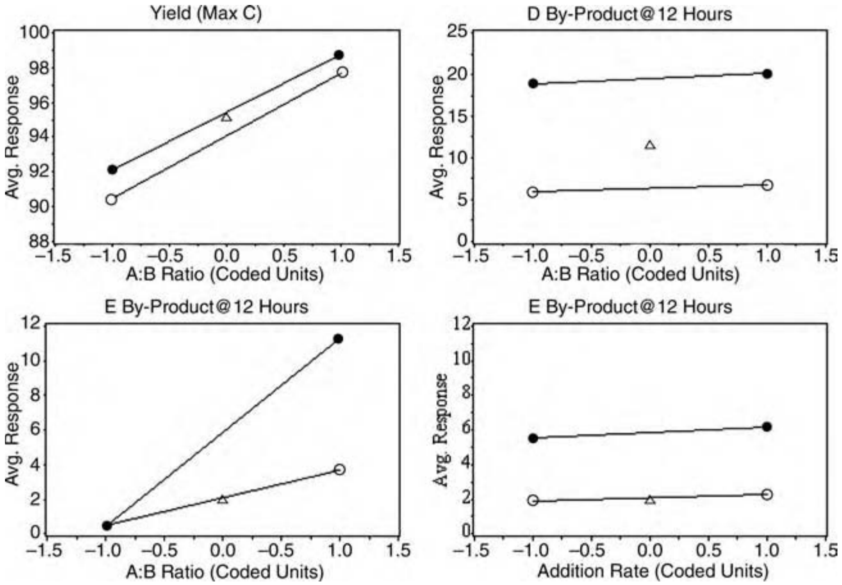


**Figure 3** Effects of plots—range-finding experiment. *Note*: high temperature = ●; low temperature = ○, mid temperature = Δ.

had a significant interaction (X1*X2) between the A:B ratio and temperature. Also, note the curvature indicated by the magnitude of the center point on the effects plot for the E by-product (lower graphs). This curvature may complicate our modeling efforts, particularly if empirical models are used. Additional design points may be necessary to build an appropriate model. But the curvature also represents an opportunity, as it suggests that there may be an operating region where we can increase yield by increasing the A:B ratio without adversely effecting the levels of the by-products.

But there is much more to be learned from this data. Now is the time to exploit the use of kinetic models. At this stage, there will be much about the kinetics that is not known, but there will be enough known that reasonable candidate models can be hypothesized. Here again the knowledge of the process chemist comes to the forefront in suggesting appropriate models to explore. From the initial studies, one can see that the rate of loss of A and B appeared to be equal and match the formation of C (see runs 1 and 5 in Fig. 1). C then appears to degrade as the reaction proceeds, particularly at higher temperatures. This suggests that a simple second-order reaction of A and B combining to form C, along with some degradation pathway for C, would be a reasonable candidate model. The formation of D and E appear to follow different pathways. The E by-product appears linked to the amount of B present (see the ANOVA results in Table 10). The D by-product appears to be linked to C. As a first step, we fit the following three models (Fig. 4).

Because we have been thorough in our sampling across time, these simple models are easily fit. This was done separately for each time profile from each experimental run. The quality of the fits was reasonable (fit diagnostics not shown), although some lack of fit was exhibited, particularly for C and E. Note that at this stage we choose not to include a degradation path for C in the model. The model will require further refining, but nevertheless the model explained much of the variation in the data. Table 11 contains the fitted rate constants.

We have taken fairly complex information and condensed it into essentially three measures—a reaction rate for the formation of C, D, and E. Not surprisingly, one thing immediately pops out—the reaction rates are very dependent on temperature (X2). We can further explore factor dependencies by applying the ANOVA technique to these rate constants. The ANOVA results (not shown) indicate that the reaction rates,

Model 1: Formation of C (no degradation pathway)

$$d[C]/dt = k_c[A][B]$$
$$d[A]/dt = -k_c[A][B]$$
$$d[B]/dt = -k_c[A][B]$$

Model 2: Formation of D

$$d[D]/dt = k_d[C]$$

Model 3: Formation of E

$$d[E]/dt = k_e[B]$$

**Figure 4** Early kinetic models.

**Table 11** Fitted Rate Constants

| | Factors | | | | | Estimated rates | | |
|---|---|---|---|---|---|---|---|---|
| Run | X1 | X2 | X3 | X4 | X5 | $k_c$ | $k_d$ | $k_e$ |
| 1 | – | – | – | – | + | 1.28 | 0.0066 | 0.0020 |
| 2 | + | – | – | – | – | 1.33 | 0.0067 | 0.0046 |
| 3 | – | + | – | – | – | 6.50 | 0.0191 | 0.0042 |
| 4 | + | + | – | – | + | 6.56 | 0.0188 | 0.0178 |
| 5 | – | – | + | – | – | 1.36 | 0.0062 | 0.0020 |
| 6 | + | – | + | – | + | 1.36 | 0.0058 | 0.0046 |
| 7 | – | + | + | – | + | 5.84 | 0.0189 | 0.0058 |
| 8 | + | + | + | – | – | 6.11 | 0.0186 | 0.0174 |
| 9 | – | – | – | + | – | 1.38 | 0.0069 | 0.0018 |
| 10 | + | – | – | + | + | 1.42 | 0.0064 | 0.0043 |
| 11 | – | + | – | + | + | 6.49 | 0.0191 | 0.0040 |
| 12 | + | + | – | + | – | 6.44 | 0.0193 | 0.0179 |
| 13 | – | – | + | + | + | 1.32 | 0.0059 | 0.0020 |
| 14 | + | – | + | + | – | 1.34 | 0.0065 | 0.0049 |
| 15 | – | + | + | + | – | 6.14 | 0.0190 | 0.0065 |
| 16 | + | + | + | + | + | 5.96 | 0.0206 | 0.0170 |
| 17 | 0 | 0 | 0 | 0 | 0 | 2.82 | 0.0106 | 0.0047 |
| 18 | 0 | 0 | 0 | 0 | 0 | 3.03 | 0.0116 | 0.0043 |

$k_c$ and $k_d$, are likely only dependent on temperature. However, the rate of formation of E, $k_e$, depends on factors X1, X2, and X3. This suggests that the models proposed for C and D are sound (except for the lack of fit observed for C noted earlier—probably due to degradation), but that the model for E may require some additional complexity. Furthermore, the values observed at the center points (see runs 17 and 18) suggest that the relationship of these rates to temperature is nonlinear. A step toward expanding the model would be to include the use of the Arrehenius relationship.

## Optimization

When the available information suggests that no further large changes are needed to meet development goals, the chemist will conduct a final round of experimentation involving the critical process factors to characterize the chemical process more accurately and to provide information for operating limits

**Table 12** Optimization Stage Summary

| Stage | Optimization |
|---|---|
| Trigger | Critical process factors and ranges identified |
| Goal | Optimization, specification, and limit setting for plant validation |
| Issues | Is the laboratory scale a good surrogate scale for manufacturing scale? Last minute process changes. |
| Approach | Study one-unit operation to minimize experimental error. Do not isolate product-measure species in process stream |
| Design | Response surface methodology |

and specification setting (Table 12). The effort now focuses on the predictability of outcomes. The emphasis is on collecting information at multiple levels of each factor in order to build and test the predictive model(s). The multiple levels allow for evaluation of the quality of the fit—are the models sufficiently flexible, where is the region of greatest change (high derivative), are there regions of poor fit?

Replication of some or all design conditions is especially critical at this stage. Replication allows us to develop sound estimates of process variation. These variance estimates serve several functions. First, they form the basis for predicting how reproducible the process will be in the plant—they predict the lot-to-lot variation. Variance estimates can also tell us about unstable regions in the reaction space. Are there regions that the plant should avoid because the process cannot be well controlled there? Finally, variance estimates will be vital to our establishment of process operating limits.

Typical design types used at this stage include central composite and Box Behnken designs. These types of designs utilize multiple levels of each factor, allowing the curvilinear or nonlinear nature of the chemistry to be explored and modeled. Fortunately, at this stage of development the experimental effort usually does not require running a complete design from scratch. Rather, the design is achieved by augmenting the previous experimental efforts. For example, a prior factorial design can be augmented with axial or star points to generate a central composite design. Generally, this augmentation will include repeats of conditions that overlap existing design points such as the center point so that any block effects that occur from one set of DOE runs to the next can be examined. Block effects are a strong indication that something critical is still unknown about the chemistry (or the analytical methods).

The analysis tools are predictive models. The goal is to be able to reliably predict outcomes within the studied factor space, describe the variation in these outcomes, and use the prediction and variance estimates to establish operating ranges. The process of building a useful predictive model is somewhat subjective but there are some useful principles.

- Err on the side of including more factors than less. Use generous inclusion criteria. For example, consider the use of alpha levels of from 0.1 to 0.2 for testing whether to keep parameters in the model.
- If an interaction is present include the corresponding main effects in the model.
- Use coded factor units if empirical linear models are to be used.
- Pay attention to model diagnostics such as residual plots.

Models can be of the linear or the kinetic type, but kinetic models should be the first choice. Kinetic models will be comprehensive at this stage, requiring simultaneous fitting of product and by-product reactions.

The development of a sound predictive model is, in the end, a kind of validation of the process. It represents a high degree of knowledge. But the model serves another goal of validation—the establishment of control limits for all critical process factors. Figure 5 illustrates the key limits and their relationship to each other.

The process for establishing limits is beyond the scope of this chapter but some key concepts are needed. The limit setting process starts with the product specifications, or what we have labeled as registration limits. These limits represent the commitment made to customers that define the product

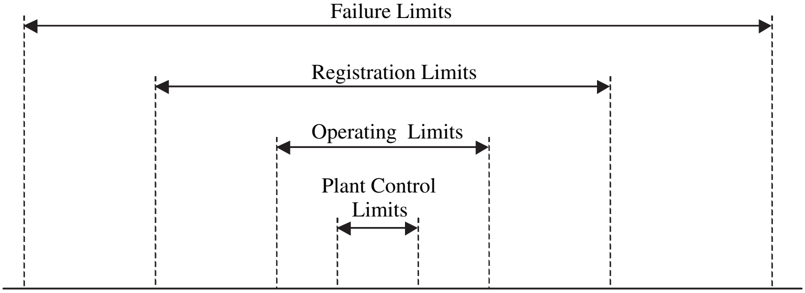


**Figure 5** Operating limits.

quality. Typically, the most important limits would be those for the type and level of chemical impurities. Registration limits are not data driven, but rather needs driven. They represent the medical science and engineering requirements of the chemistry. So for example if the "customer" is the next step in the manufacturing process then an impurity specification would depend on what the next downstream process can handle. If the customer is the consumer then the specification might be defind by the toxicity profile of the impurity.

Operating limits are data driven, with the most critical measure being the process variation. They are established by working inward from the registration limits to provide assurance that we will not fail registration limits because of process or assay noise. At this stage of development the process variation will not be completely known, particularly in the hands of the plant, but the noise observed from the DOEs can be a reasonable starting point.

Two other limits are given in the figure. The plant control limits are the usual $3\sigma$ control limits. They are a trouble detection mechanism designed to distinguish common cause variation from special cause. The failure limits represent the point at which product can no longer be successfully processed downstream. Note that knowing how close the operating limits are to the failure limits, defines whether the factor is a critical process factor. The relationship illustrated in Figure 5 is the ideal–we may not always be so fortunate. For example the plant control limits could lie outside of our operating limits. If this is the case then the process could not be said to be validated since there is too high a likelihood of failing the specifications.

The predictive models that we have taken pains to establish allow us to translate these product limits (in the response or “y” dimension) into factor ranges (the “x” dimension). The ranges need to account not only for the process variation but also the uncertainty in the models themselves. The factor ranges also must be within the manufacturing site’s ability to control otherwise additional noise will be introduced. For example, a successful manufacturing process cannot require temperature control within 2°C if the plant control capabilities are 5°C. These ranges can be a challenge to deliver, but the ability to describe an operating range and accurately predict outcomes within that range represents a high degree of achievement. Ultimately it is validation.

## Example

In the last round of experiments we established two key factors: the ratio A:B, and temperature. A third factor of lesser importance is the rate of addition. Recall that the design used was a $2^{5-1}$ with duplicated center points. Because only three factors seemed to be contributing, the remaining factors were dropped and the results collapsed into a full factorial design. Figure 6 illustrates the collapsed design space for the observed $C_{\max}$ response from Table 9. Notice that we end up with replicate data for all the factor combinations. We have gained a great deal of information about process variation. In this example, the variance appears to be reasonably consistent across the design space. We can also begin to better visualize how the reaction responds in the factor space.

The team decided to augment the previous design with “star points” and additional center points to create a central composite design. This should improve our ability to test the quality of the fitted model, particularly the degree of curvature if polynomials are used. The center points were included to test for any block effects. Table 13 lists the augmented design points along with the resulting key responses (once again the time profiles are too detailed to be included here). The design levels are given in coded units. Notice that for two of the design

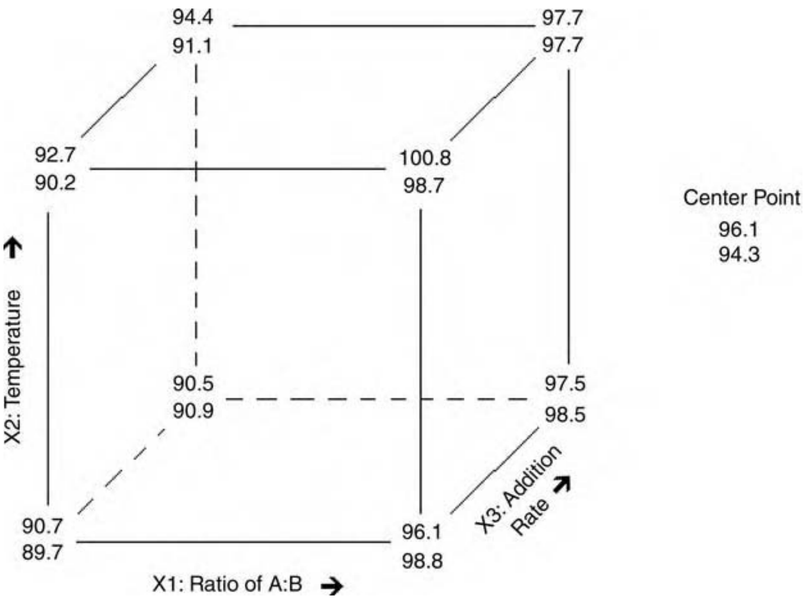


**Figure 6** Results from range finding.

runs the factor levels have been modified from what would be typical levels (±1.7) for star points in a three-factor design. The first run increased the level for X1 (A:B ratio) from –1.7 to –1.3. This corresponds to a ratio of 1:0.9 rather than 1:0.8. It was felt that shorting the reaction too much would not provide useful data. Also for run 6 the level for X3 is +1.0 as this already represents the maximum possible addition rate for B. These

**Table 13** Axial Points with Center Point

| Run | X1 | X2 | X3 | $C_{max}$ (%) | D-12 hr (%) | E-12 hr (%) |
|---|---|---|---|---|---|---|
| 1 | –1.3 | 0 | 0 | 82.3 | 9.7 | 0.3 |
| 2 | +1.7 | 0 | 0 | 104.8 | 11.7 | 10.2 |
| 3 | 0 | –1.7 | 0 | 93.0 | 4.1 | 1.3 |
| 4 | 0 | +1.7 | 0 | 96.6 | 36.8 | 6.7 |
| 5 | 0 | 0 | –1.7 | 94.6 | 11.2 | 1.9 |
| 6 | 0 | 0 | +1.0 | 91.9 | 11.0 | 2.3 |
| 7 | 0 | 0 | 0 | 93.8 | 11.5 | 2.0 |
| 8 | 0 | 0 | 0 | 96.7 | 11.1 | 2.2 |

kinds of compromises away from standard levels are often necessary in central composite designs.

The results indicate a well-behaved system. There were no failures. The center points compare well with what was obtained from the range-finding experiment and the other results compare well with expectations. The analysis combined these data points along with those from the range-finding design (the screening data could be included as well) into a single comprehensive data set. A predictive model was then fit. As noted, the model could be based on linear polynomials or kinetic models. We will illustrate the latter.

Based on everything learned to date, a kinetic model was proposed (Fig. 7). Note the interconnectedness of the model—different responses share common parameters. This will require the fitting of both parent and by-products simultaneously. In addition, it was proposed to include temperature in the model using the Arrhenius relationship. Thus, each rate constant was considered to be of the form:

$$k = Fe^{-\mathrm{E}_{\mathrm{A}}/T} \tag{1}$$

where $F$ is a frequency factor parameter, $\mathrm{E}_{\mathrm{A}}$ is a parameter that depends on the ratio of the activation energy to the molar gas constant, and $T$ is temperature (°K).

Note that models are never completely comprehensive. Compare this model with the true model assumed in the appendix. Complete understanding is elusive, but that does not detract from the utility of less than perfect models.

The model was fit, and the quality of fit was determined to be good. The fitted model can be visualized as three response

Kinetic Model:

$$\begin{aligned}
d[C]/dt &= k_1[A][B] - k_2[C] \\
d[A]/dt &= -k_1[A][B] \\
d[B]/dt &= -k_1[A][B] - k_3[B] \\
d[D]/dt &= k_2[C] \\
d[E]/dt &= k_3[B]
\end{aligned}$$

**Figure 7** Final kinetic model.

surfaces, one for C, D, and E, each in four factor dimensions—the ratio A:B, the addition rate of B, temperature, and time. An immediate use of the model is to find combinations of factor levels that meet our goal of maximizing yield while maintaining-the levels of D and E below 1%. A search of the fitted space reveals that the "best" operating range is at high temperature and a very rapid (batch mode) addition rate of B. Figure 8 presents the fitted contours for yield and by-products for the batch addition mode (rate = 1000 units/hr) at a temperature of 50°.

From the contour plot one can see that to achieve levels of D below 1% requires keeping the reaction time under approximately 0.3 hours (20 minutes) when the reaction is run at 50°. Furthermore, if the reaction is run for at least 0.2 hours, then yields in the neighborhood of 95% to 98% are possible as long as an excess amount of B is added (approximately 1.8 units or greater).

These results do not yet represent an operating range. We need to use what we know about the noise in the process

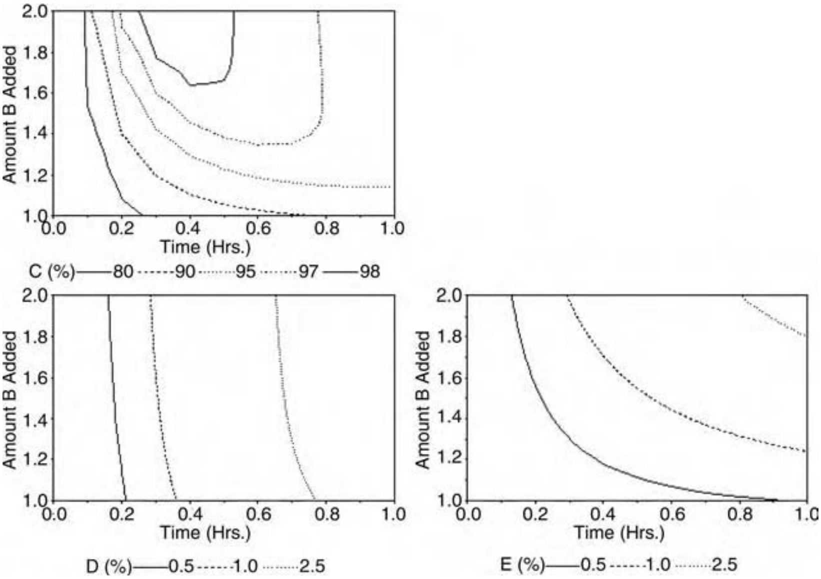


**Figure 8** Fitted contour plots for percent C, D (temp = 50° and batch addition mode).

to adjust our approximate optimal settings to account for randomness in the next batch(es). As we have noted, the formal process of accounting for process variation is beyond the scope of this chapter, but we have all the information required to establish operating ranges that are supported by data.

## SUMMARY

We have taken pains to lay out process validation as a methodical, sequential process. We outlined a flexible sequence of experimental designs that will gather information quickly and efficiently. The importance of upfront planning and working with subject matter experts for factor and range selection was stressed. We demonstrated the notion of expanding the number of sample points during each run of the experiment to achieve a deeper understanding of the chemical reaction. These extra data points permitted the construction of concentration–time plots which are the backbone of understanding of chemical reactions. From these plots, reaction mechanisms can be postulated and rate laws can be developed to support model building. The ultimate outcome of a validated chemical reaction is achieved, critical process parameters are identified and characterized.

## REFERENCES

1. Department of Health and Human Services, U.S. Food and Drug Administration. Pharmaceutical cGMP's for the 21st Century—A Risk-Based Approach. Final Report, Fall 2004.
2. Box GEP, Hunter WG, Hunter JS. Statistics for Experimenters. New York: John Wiley & Sons, 1978.
3. Coleman DE, Montgomery DC. A systematic approach to planning for a designed industrial experiment. Technometrics 1993; 35:1–12.
4. Box GEP, Youle PV. The exploration and exploitation of response surfaces: an example of the link between the fitted surface and the basic mechanism of the system. Biometrics 1955; 287–323.

# Appendix 1

The example used to illustrate the design and analysis concept is based on a simulated chemical reaction. The use of simulated rather than real data allows for better illustration of the entire design process. The reaction assumed was simple: a starting material, A, is combined with a reagent, B, to form product, C. By-products D and E are formed from both C and B, respectively. The true mechanism was assumed to follow the kinetic model described by the following differential equations,

$$
\begin{aligned}
&d[C]/dt = k_1[A][B] - k_2[C] \\
&d[A]/dt = -k_1[A][B] \\
&d[B]/dt = -k_1[A][B] - k_3[B][B] \\
&d[D]/dt = -k_2[C] \\
&d[E]/dt = -k_3[B][B]
\end{aligned}
$$

We further assume that the rates of reaction can be described by the Arrhenius equation,

$$k_L = Fe^{-\mathrm{E_A}/T}$$

where $T$ is temperature in Kelvin, and $F$ and $\mathrm{E_A}$ are parameters representing a frequency factor and the ratio of the activation energy to the molar gas constant, respectively.

To facilitate learning, we have chosen a very clean chemistry with only one reaction step. Other factors, particularly those that are more physical in nature were assumed to have no effect. So, for example, agitation levels, mixing times, starting material purity, pH, and so on would not affect the reaction. It is of course assumed that this information is unknown when starting the experimental effort. These factors would most certainly be included early in the design discussions.

For each design, data were generated using the theoretical kinetic model, the factor levels specified in the design, and a random noise component. Without loss of generality it was assumed that $[A] = 1$.

3

# The Role of Designed Experiments in Validation and Process Analytical Technologies

**LYNN D. TORBECK**

Torbeck & Associates, Inc.
Evanston, Illinois, U.S.A.

**RONALD C. BRANNING**

Genentech, Inc.
South San Francisco, California, U.S.A.

## INTRODUCTION

Validation and process analytical technology (PAT) require that we find and control the primary sources of product and process variation. Not many people realize that the current good manufacturing practice contain the words validation and variability in the same sentence. "Such control procedures shall be established to monitor the output and to validate the performance of those manufacturing processes that may be responsible for causing variability in the characteristics of in-process material and the drug product" (1).

To accomplish this, we must collect data. But how we collect that data is as important as the data themselves. Some data are worthless, some are priceless. The conditions and procedures used to find data ultimately determine their value. Statistical quality control (SQC), statistical process control (SPC), total quality management (TQM), and six sigma are all passive approaches to data collection. These procedures only observe and report what is happening. They cannot find the analytical cause-and-effect relationships needed for true process understanding and for controlling the sources of variability.

Controlled multivariate experiments are the most logical, the most scientific, and the most efficient way that scientists know to collect data. Controlled experiments are the scientific

The basis of this chapter is from a paper by the same authors titled "Designed Experiments—A Vital Role in Validation," published in *Pharmaceutical Technology*, June 1996.

method in action and applied at the laboratory bench and on the production floor. As a result, experiments have an increasingly vital role to play in validation, PAT, and thus the quality of the products. In 1985, E. M. Fry, then director of Food and Drug Administration (FDA)'s Division of Drug Quality Compliance, announced that data were critically important in validation (2).

> Validation has a quantitative aspect—it's not just that you demonstrate that a process does what it purports to do; you actually have to measure how well it does that. Then, the processes that cause that variability must be identified. Experiments are conducted (that is, validation runs) to ensure that factors that would cause variability, are under control.
>
> The regulations require validation of those processes responsible for causing variabilities in the characteristics of in-process materials or finished products. However, the regulation implies that not everything that takes place in a pharmaceutical manufacturing plant causes variability. Therefore, some things don't have to be validated. We never intended to require that everything [that] takes place in a manufacturing operation is subject to a validation study.

Recall the FDA definition of validation: "Validation is establishing documented evidence which provides a high degree of assurance that a specific process will consistently produce a product meeting its predetermined specifications and quality attributes" (3). Clearly, consistently means a lack of variability. But how should variability be reduced? In the "Guideline on General Principles of Process Validation," FDA states, "Quality, safety, and effectiveness must be designed and built into the product; quality cannot be inspected or tested into the finished product" (3). Thus, minimizing variation must start in development.

K. G. Chapman observed that

> Process development equals process validation plus process optimization. A well-developed process is, therefore, by definition a well-validated process. Once it is decided that a bulk pharmaceutical chemical (BPC) process should be validated, the question becomes "How?" In the case of a new process, the answer is simple: Do a good process development job and document it (4).

He also presents a validation timeline that includes a product's life cycle, including the design stage.

Note that nowhere in Fry's discussion, FDA's definition, or Chapman's validation timeline are we told where or when the data needs to be collected for validation. Although the FDA definition specifies "predetermined," obviously the data used to set the specification are not useless later.

Supporting this, H. L. Avallone notes that

> Process validation for a bulk pharmaceutical chemical (BPC) may include development data that describe the limitations and efficiency of the process … Their laboratory notebooks and their processing records also may be reviewed because these records may constitute the raw data for process validation (5).

We must collect validation data throughout the product and process development life cycle. To demand that all validation data be collected only after development is completed assumes that nothing of importance was learned during development. On the contrary, in some fortunate projects, validation is nearly completed when development is completed. Simplicity in validation can be realized.

Fry continues, "We are saying a process should not be operated under worst case conditions which have not been included in validation studies. To put it another way, don't operate a process in uncharted waters" (2). Anyone who has gone boating will identify with this philosophy: If you don't want your keel ripped off, you must know how the bottom of the ocean looks.

> Effective March 12, 2004, FDA revised a long-standing policy document regarding the validation of pharmaceutical manufacturing processes for drugs that are subject to pre-market approval requirements. This policy guide is now titled *Process Validation Requirements for Drug Products and Active Pharmaceutical Ingredients Subject to Pre-Market Approval* …. New to this version is the recognition of the role of emerging advanced engineering principles and control technologies in ensuring batch quality…. This version also deletes the previous reference to "three" validation (or conformance) batches at commercial scale as adequate proof of process validity—a number is no longer suggested.

> The proof of validation is obtained through rational experimental design and evaluation of data, preferably beginning from the process development phase and continuing through the commercial production phase. Prior to the manufacture of the conformance batches, the manufacturer should have identified and controlled all critical sources of variability (6).

In September 2004, the FDA issued the Guidance Industry, *PAT—A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance* (7).

> This guidance is intended to describe a regulatory framework process analytical technology (PAT) that will encourage the voluntary development and implementation of innovative pharmaceutical development, manufacturing, and quality assurance.... Our new strategy is intended to alleviate concern among manufacturers that innovation in manufacturing and quality assurance will result in regulatory impasse.... The goal of PAT is to enhance understanding and control the manufacturing process, which is consistent with our current drug quality system: *quality cannot be tested into products; it should be built-in or should be by design*.... This benefit can be achieved through the use of multivariate mathematical approaches, such as statistical design of experiments (DOE), response surface methodologies,.... Methodological experiments based on statistical principles of orthogonality, reference distribution, and randomization, provide effective means for identifying and studying the effect and interaction of product and process variables.... Develop mathematical relationships between product quality attributes and measurements of critical material and process attributes.

Thus, it is clear that the FDA has promoted a consistent message since 1978: find and control the sources of variation using good science and good statistical practice. To do that, we need to collect data.

## DATA COLLECTION

How should the vast oceans of a manufacturing process be mapped? How should factors that cause variability be separated from those that do not? How should data for validation be collected in the most scientific and cost-effective way?

Scientists and engineers collect data either by observation or by experiment. In the observational approach, the investigator is only an observer of the product or the process. The investigator records observations in the manner of Charles Darwin sailing on the ship the Beagle or Jane Goodall in Africa. Both recorded their observational landmark studies; they made no attempt to control the environment or influence events.

However, cause-and-effect relationships in these situations are obscured by rampant variability and multiple mysterious causes. The approach is passive. Classical observational tools for industry usually include sampling plans, control charts, and process capability studies. In addition, Branning has found two of the most useful observational tools for validation and PAT are process flow charts and fishbone diagrams, which help define the process and identify the potential sources of variability. These observational tools need to be used on a routine basis to collect background data for validation and PAT.

However, as noted, these tools are passive. There is no deliberate and specific control of the environment or critical process parameters. These observational tools cannot find and describe cause-and-effect relationships directly. The only way to find these relationships is to conduct a multivariate controlled experiment.

In contrast to the observational approach, data collection in a controlled experiment is active; investigators take control of the environment and critical process parameters. By deliberate changes in key factors, the cause-and-effect relationships are forced to show themselves.

## CONTROLLED EXPERIMENTS

There are three ways scientists and engineers conduct controlled experiments: success/failure, one-factor-at-a-time, and multiple factors at a time. Humans have used the first two since before recorded history, and still use them intuitively in many facets of our lives. In 1843, John S. Mill described these two approaches in his book *Systems of Logic*. He called the success/failure experiment the absolute experiment, and the one-factor-at-a-time experiment the method of differences. Both have become ingrained and implicit in western science. Unfortunately, one author has specifically called for the one-factor-at-a-time approach for validation.

> All process variables should be monitored, and never more than one varied at a time so that the effects can be fairly and accurately evaluated.... Then we can challenge the process by intentionally changing adjustments to affect only one process variable at a time to reach and then exceed the acceptance upper/lower limits of each specification (8).

Although this is appealing, it is grossly inefficient. There is a third and more effective scientific method for collecting data experimentally. Sir R. A. Fisher, a geneticist and mathematician, originated the multiple-factors-at-a-time approach. He developed this advance at Rothamsted Agricultural Research Station when working in Harpenden, England, starting in 1919. He wrote the first journal article in 1926 (9), followed in 1935 by a textbook, *The Design of Experiments*, which is still in print (10). The first books showing the approach's industrial applications were published in the early 1950s (11–13).

The multiple-factor approach has been shown in practice to use resources more efficiently, and many scientific fields and industries use DOE extensively. But this approach is still unknown to some areas as the basic concepts are not routinely taught to undergraduate science and engineering students.

The use of DOE in the pharmaceutical industry began early. For example, S. M. Free, while head of statistics at Smith Kline & French in 1957, wrote an internal statistics booklet with F. A. Oyer (14). It showed how DOE could be used in formulation development and stability studies. Now called matrixing, this approach is now referenced in International Conference Harmonisation (ICH) Guideline Q1 (15). Many articles on DOE have appeared in the pharmaceutical literature during the past 25 years (16–31). Currently, the routine use of DOE is sporadic. The key to its success is an advocate in management who provides the needed motivation and incentives.

Using DOE for pharmaceutical validation has historically been less well known. Torbeck has been giving talks and papers since the late 1970s showing the application of designed experiments to validation and conducts training in the subject (32,33). Both authors have given numerous presentations about using DOE for process validation. To the authors' knowledge, the first published application of designed experiments for validation may be by A. Y. Chao et al. (34).

In the context of validation and PAT, designed experiments are used throughout the development journey to find and quantify cause-and-effect relationships. The data collected to develop the product and its specifications then also support validation. Concurrent development and validation data collection are realized. Thus, the factors causing variability are identified, quantified, and controlled. Every point on Chapman's validation timeline benefits from DOE data collection, which is used most effectively in the design stage and in prospective performance qualification (4). The concept of a timeline is in the PAT guidance and in the GMP revision as a lifecycle.

Brainstorming sessions for process flow charts typically name 25 to 30 factors believed to have some effect on the outcome. However, the Pareto concept of the vital few and trivial many, also called the 80/20 rule, indicates that usually not more than seven, and often five or less, factors are truly significant and vitally important. Chapman notes this as well: "A [validation certificate] is promulgated for each operating parameter range designated as critical in the control spreadsheet; most pharmaceutical processes require about five to 10 certificates" (35).

We typically use two levels—high and low—to conduct initial experiments. This correlates with Chapman's proven acceptable range (PAR) approach (35): "Each end of each PAR must be supported by documented evidence; otherwise, is not a proven range." However, some authors believe that looking at five to 10 factors each at two levels requires them to study all possible combinations of factors and levels. For example, statements in Berry (36) are extended by Sharp (37):

> Let us assume a simple product with only three ingredients, each with five test parameters such as those suggested. That aspect of the "process challenge" will require the manufacture of (25)3 = 32,768 experimental batches. Logically, they would need to be combined with all possible combinations of the "process variable batches" already discussed, and we would then have a total of 35 × (25)3 = 7,962,624 experimental batches. They would, of course, have to be full-scale production batches (37).

This is manifestly misleading! Thousands of journal articles and hundreds of books (11–13) on designed experiments, many of them pharmaceutical (16–31,38,39), clearly show that three to 15 factors, in as few as 8, 12, or 16 sets of conditions, in small-scale development batches, can be efficiently studied. These are not full-scale production batches. If done early in development, the DOE data used to develop the product can serve to set specifications and to document validation. This is already being done in some companies for assay validation (40). "Experiments conducted during product and process development can serve as building blocks of knowledge that grow to accommodate a higher degree of complexity throughout the life of a product" (7). From a managerial perspective, Branning has found the use of DOE results in productivity improvements ranging from 25% to 200% and has reduced project time and costs by 25% to 50%.

Design of experiments not only finds factors in complex processes that have a significant impact in their own right, but it can also find the joint interaction effects between these factors. The observational, success/failure, or one-factor-at-a-time approaches simply cannot find these interactions. When a process is described as more art than science, this usually suggests the joint interactions of factors that are not understood. The fact that those interactions are important is reinforced by the PMA Validation Advisory Committee:

> Consideration must also be given to the potential interactive influence of other parameters in the total system. Adverse effects of an extremely high or low pH, for example, might be aggravated by extending time and/or elevated temperature (41).

"Traditional one-factor-at-a-time experiments do not address interactions among product and process variables" (7).

Another major benefit that is often overlooked is the identification of factors that do not affect process or variability. These factors can then be set to their most economical level, or specifications may be relaxed. Data and statistical analysis collected during the experiment will support these actions.

## CONCLUSION

Validation and PAT need effectiveness (validating the right things) and efficiency (validating the best way). Given the time, money, and staffing constraints imposed by the marketplace, more time must be spent on planning for better execution. This is best accomplished by using all available tools. Especially needed are basic observational tools and experimental methods, including the success/failure and one-factor-at-a-time approaches. But multifactor-at-a-time experiments, DOE, also have a vital and substantial role to play in developing process knowledge for validation.

They are the most cost-effective way to collect data to study cause-and-effect relationships. The use of DOE is needed to gain efficiency, study joint interaction effects, and identify factors controlling variability. Today's competitive business and regulatory environments demand that the best approaches be learned and used.

## REFERENCES

1. FDA (1978, revised 2001). Current Good Manufacturing Practice, 21 CFR 211. 110(a).
2. Fry EM. The FDA's viewpoint. Drug Cosmetic Ind 1985; 137(1):46–51.
3. FDA, Guideline on General Principles of Process Validation (1987). http://www.fda.gov/cder/guidance/pv.htm.
4. Chapman KG. A history of validation in the United States: Part I. Pharm Technol 1991; 15(10):82–96.

5. Avallone HL. GMP inspections of drug-substance manufacturers. Pharm Technol 1992; 16(6):46–53.

6. FDA, Compliance Policy Guide 7132c.08, Process Validation Requirements for Drug Products and Active pharmaceutical Ingredients, March 12, 2004.

7. FDA(2004). PAT—A Framework for Innov ative Pharmaceutical Development, Manufacturing, and Quality Assurance.

8. Berry IR. Practical process validation of pharmaceutical products. Drug Cosmetic Ind 1986; 139(3):36–46.

9. Fisher RA. The arrangement of field experiments. J Ministry Agric Engl 1926; 33:503–513.

10. Fisher RA. The Design of Experiments. Edinburg, U.K.: Oliver and Boyd, 1935.

11. Tippett LHC. Technological Applications of Statistics. New York, NY: John Wiley & Sons, 1950.

12. Davies OL. The Design and Analysis of Industrial Experiments. London: Longman Group Limited, 1954.

13. Bennett CA, Franklin NL. Statistical Analysis in Chemistry and Chemical Industry. New York, NY: John Wiley & Sons, 1954.

14. Free SM, Oyer FA. "Statistical Guides to Pharmaceutical Formulation," internal document, Smith Kline & French, June 1957.

15. ICH Q1A(R2), Stability Testing of New Drug Substances and Products, November 2003.

16. Bolyard KB, McCurdy VE. Evaluation of a cartridge and a bag filter system in fluid-bed drying. Pharm Technol 1994; 18(3):104–114.

17. Gohel MC, Patel LD, Modi CJ, Jogani PD. Functionality testing of a coprocessed diluent containing lactose and microcrystalline cellulose. Pharm Technol Yearbook 1999; 40–46.

18. Gohel MC, Jogani PD. An investigation of the direct-compression characteristics of coprocessed lactose-microcrystalline cellulose using statistical design. Pharm Technol, Yearbook, 1999; 54–62.

19. Walsh A, Smith S, Stanley J, et al. Development and validation of automated methods for finished product testing. Pharm Technol March 2000; 134–148.

20. Konkel P, Mielck JB. A compaction study of directly compressible vitamin preparations for the development of a chewable tablet: Part I. Pharm Technol 1992; 138.

21. Konkel P, Mielck JB. A compaction study of directly compressible vitamin preparations for the development of a chewable tablet: Part II. Pharm Technol 1992; 42.

22. Wehrle P, Colombo P, Santi P, et al. response surface methodology applied to fluid bed granulation. Pharm Technol Eur April 2000; 36–46.

23. Hartauer KJ, Bucko JH, Cooke GG, et al. The effect of rayon coiler on the dissolution stability of hard-shell gelatin capsules. Pharm Technol 1993; 17(6):76–83.

24. Kelly BD, Jennings P, Wright R, Briasco C. Demonstrating process robustness for chromatographic purification of a recombinant protein. BioPharm, October 1997.

25. Price J. Blow-fill-seal technology: Part II, design optimization of a particulate control system. Pharm Technol February 1999.

26. Peters PA, Paino TC. Robustness testing of an HPLC method using experimental design. Pharm Technol, Analytical Validation 1999.

27. Hwang RC, Gemoules MK, Ramlose DS, Thomasson CE. A systematic formulation optimization process for a generic pharmaceutical tablet. Pharm Technol May 1998.

28. Dobberstein RH, Corkle WJ, Million G, et al. Computer-assisted experimental design in pharmaceutical formulation. Pharm Technol 1994; 18(3):84–94.

29. Bandurek GR. Using designed experiments in validation. BioPharm Int May 2005; 40–52.

30. Murphy MP, Hollenbeck RG. A unique application of extrusion for the preparation of water-soluble tablets. Pharm Technol April 1998; 94–104.

31. Wiseman D, Griffiths KL. Implementing statistically designed experiments in analytical method validation. Am Pharm Outsourcing September/October 2004; 13–25.

32. Torbeck LD. Roundtable Discussion of Statistics and GMPs. Paper presented at the First Annual Midwest Biopharmaceutical Statistics Workshop, Ball State University, Muncie, IN, May 1978.

33. Torbeck LD. Validation by Design®, short course. Evanston, IL: Torbeck and Associates.

34. Chao AY, Forbes ES, Johnson RE, von Doehren P. Prospective process validation. In: Loftus BT, Nash RA, eds. Pharmaceutical Process Validation. New York: Marcel Dekker, 1984:125–148.

35. Chapman KG. The PAR approach to process validation. Pharm Technol 1984; 8(12):22–36.

36. Berry IR. Process validation of raw materials. In: Loftus BT, Nash RA, eds. Pharmaceutical Process Validation. New York: Marcel Dekker, 1984:203–249.

37. Sharp JR. The problems of process validation. Pharm J 1986; (1):43–45.

38. Lewis GA, Mathieu D, Phan-Tan-Luu R. Pharmaceutical Experimental Design. New York, NY: Marcel Dekker, 1999.

39. Bolton S. Pharmaceutical Statistics, 2nd ed. New York, NY: Marcel Dekker, 1990.

40. Torbeck LD. Assay validation: ruggedness and robustness with designed experiments. Pharm Technol 1996; 20(3):168–172.

41. PMA Validation Advisory Committee. Process validation concepts for drug products. Pharm Technol 1985; 9(9):78–82.

# 4

# Experimental Design for Bioassay Development and Validation

**DAVID M. LANSKY**

Lansky Consulting, LLC, d/b/a Precision Bioassay
Burlington, Vermont, U.S.A.

## INTRODUCTION

Biological assays using cultured cells are a crucial part of the quality-control process for pharmaceutical protein products. These assays can have large variation and can require substantial effort to develop, perform, analyze, and validate. The assignment of reference samples and dilutions to locations on 96-well cell-culture plates is only one aspect of experimental design for these assays. This chapter will suggest how classical experimental design methods (factorial screening designs, response-surface designs, Latin Square, split-plot, strip-plot, and nested designs) can be effectively used in the development, routine performance, and validation of cell-culture bioassays.

Biological assays are systems that use comparisons among groups of living organisms to measure a response. These assays are typically used to measure proteins or other complex molecules that have both very specific activity and very high potency. Because biological assays report results based on a comparison among groups of organisms, the concept of relative potency is fundamental in these assays. For each sample (where a reference or standard is also considered a sample), a concentration–response relationship is demonstrated experimentally. Assuming that the groups of organisms are comparable and the concentration–response relationship is similar, the difference between the concentration–response relationships is interpreted as a relative potency. Note that much depends on the exact definition of similar. The typical assumptions of biological assay include:

- samples contain the exact same active compound, differing only in the amount or concentration of analyte;
- the concentration–response relationship has constant variance around the fitted model (perhaps after a transformation of the response);
- the variation around the fitted model is normally distributed;
- observations are independent (1–3).

The constant variance assumption can be relaxed via either a rescaling of the response or a weighted fit (4). Similarly, if an appropriate model is used, the normality assumption may be relaxed (4). For example, with a dichotomous response, a logit-log model may be appropriate (5). Other response patterns (e.g., Poisson) may be fit via a generalized linear model (6). For quantitative responses, it is often most practical to find a rescaling or transformation of the response scale to achieve nearly constant variance and nearly normal responses. Finally, if samples are grouped, then blocks or other experiment design structures must be included in the model (7–12).

This chapter does not focus on the methods for assessing similarity or parallelism; for background, see Plikaytis et al. (13). A newer and much better approach to assessing parallelism uses confidence intervals (14,15). The focus here is on how to use experimental design to accelerate the assay-development process; methods for addressing concerns about potential correlations among groups of observations; and how practical constraints influence assay design. A major goal is to show an effective way to perform and summarize validation experiments for biological assays (16) that allows flexible use of the assay for those who must set assay and product specifications.

Most assays can be thought of as having a concentration–response curve that is approximately logistic in shape (1,13); that is, for low concentrations of analyte, the assay response is consistently low (or high). For some (typically quite narrow) range of concentrations of analyte, the response increases

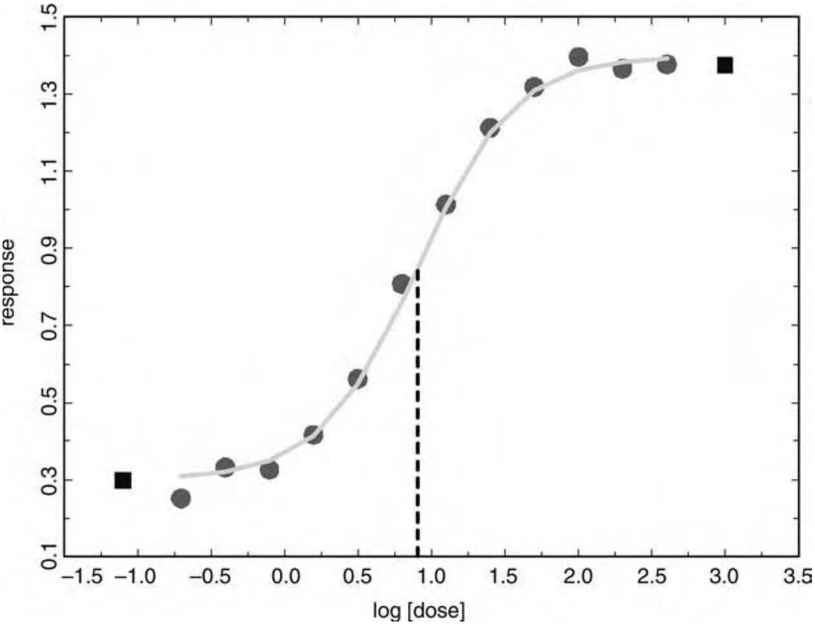


**Figure 1** A typical logistic log concentration–response curve showing the true (unknowable) curve as the line, round points, with random normal variation around the curve, and both positive and negative controls as square points (the controls are plotted using artificially imposed dose levels to make the plot easy to interpret). For this example, the lower asymptote is consistent with the low control and the upper asymptote is consistent with the high control; in many assay systems, one or both of these controls are not consistent with the asymptotes. The dashed vertical line indicates the EC50 or ED50, the log concentration that gives a response midway between the two asymptotes.

(or decreases) rapidly with increasing concentration of analyte. For high concentrations of analyte, the response is consistently high (or low) (Fig. 1). The low- and high-response regions are called the asymptotes of the response, whereas the steep portion of the response curve is called the responsive region.

## PRELIMINARY DEVELOPMENT EXPERIMENTS

The goals during preliminary development include finding a set of assay conditions where the assay is responsive to the analyte(s) of interest and where the assay response is not sensitive to other factors that may be either controlled or uncontrolled (16–18). The desirable properties of an assay concentration–response curve include having good separation between the asymptotes and having small variation around the response curve. A very steep response curve is sometimes inconvenient; in particular, with a steep response curve and asymptotes that are not well separated, it is difficult to get several responses on the steep part of the response curve, especially if the EC50 of the curve (the dose of a sample that yields a response halfway between the asymptotes of the dose–response curve) varies appreciably from day to day, requiring the use of a wide range of concentrations to ensure capturing both asymptotes. Even so, it is easier to get good precision on the relative potency from a very steep curve than from a very shallow response curve. A particularly useful summary of a response curve in early development is to divide the difference in the asymptotes by the estimated standard deviation around the curve; we seek conditions that will maximize the ratio,

$$t = \frac{\text{upper asymptote} - \text{lower asymptote}}{S_{\text{pooled}}}$$

Note that this ratio is meaningless unless the variation around the curve is constant throughout the concentration and response range; hence an appropriate transformation must be applied before this summary.

In initial range-finding experiments, it is often practical to do one-factor-at-a-time experiments (4), followed by factorial screening designs using many factors (4,7), each at two levels. As we are interested in the properties of the concentration–response curve (particularly the difference between the asymptotes, the steepness of the responsive region, and the variation around the curve), it is important to quickly move to study the full curve. For cell-culture assays performed in 96-well plates, an effective approach is to assign concentrations of the analyte to the columns. Alternatively, unique combinations of all other treatment factors are assigned to rows, plates, or runs. Runs are groups of plates set up at one time by one analyst using a common batch of cells. During development, it is important to identify large location effects and large sources of variation while using small numbers of replicates of many different operating conditions. It is often very useful to study the full concentration–response curve at each condition. A powerful and convenient technique is to assign individual rows of assay plates to unique combinations of conditions in factorial or response-surface designs for factors that may affect the assay while assigning concentrations (dilutions) to columns. A few two-level factorial (or fractional factorial) designs for factors other than analyte concentration is a very effective way to find good initial operating conditions. Careful attention to experimental units is important here as some factors can be assigned to plate rows (e.g., buffer and cell concentration), whereas others must be assigned to entire plates (e.g., the length of the incubation time and the number of washes). Others are most appropriately assigned to groups of plates (e.g., properties of the preassay cell-culture conditions and incubation temperature) (11,12,19,20).

After having identified good operating regions, it is useful to use a response-surface design (4,10) to find good conditions for combinations of factors that can be set quantitatively. For example, in an antiviral assay, it is likely that the starting cell density and the starting concentration of virus would need to be carefully optimized along with the preassay cell-culture conditions. At this stage (if not earlier), it is also appropriate to begin collecting additional information about sources of variation. Factors that may be of interest at this stage include analysts, batches of cells, plates, and rows or columns within the plate.

## PLATE LAYOUT AND ASSAY UNIT: STRATEGIC DECISIONS

Assays with moderate to large variation in the parameters of the response curve associated with either plates or sections of plates are likely to benefit from plate layouts and analysis methods that protect the assay results from location effects. Rather than use absolute measures of variation in these parameters as guidance, it is more appropriate to ask how these sources of variation compare with other measured sources of variation. For example, if the row-to-row variation in curve parameters is an order of magnitude smaller than the pooled estimate of variation around curves fit to each row, then it may not be necessary to use design and analysis methods to protect the assay against row effects.

Even without evidence that location effects are a concern, consideration of experimental units (details in the next paragraph) may lead to assay designs (and associated analyses) that will provide good protection against location effects. In most assay laboratories, samples are assigned to plate rows (or columns) and sample concentrations are assigned to plate columns (or rows). Often a multichannel pipette is used to dilute multiple samples simultaneously.

By definition, the experimental unit is the smallest unit randomly allocated to a distinct level of a treatment factor. Note that if there is no randomization, there is no experimental unit and (in nearly all cases) no experiment. Although it is possible to perform experiments without randomization, it is difficult to do well, and risky unless the experimental system is very well understood (7). Randomization is important for several reasons. Randomization changes the sources of bias into sources of variation; in general, a noisy assay is better than a biased assay. Further, randomization allows estimates of variation to represent variation in the population; this in turn justifies statistical inference (standard errors, confidence intervals, etc.). A common practice in cell-culture bioassay is to rotate among a small collection of layouts rather than use random allocation. Whereas rotation among a collection of layouts is certainly better than a fixed layout, it is both possible and practical to use carefully structured randomization on a routine basis, particularly when using a robot.

When samples are assigned to, for example, rows, the row becomes the experimental unit for the sample. Similarly, when analyte concentrations are assigned to columns, the column becomes the experimental unit for concentration. This type of layout should be analyzed as a strip-plot or split-block design where the blocks may be plates or sections of plates (8–12). A further step would be to balance the locations of samples across rows and concentrations across columns, possibly using a double Latin square strip-plot design as detailed in Lansky (12).

Even in the absence of substantial data demonstrating that location effects are not a concern, it is wise to use a design that will both protect the assay results from location effects and allow the monitoring of location effects using the data produced by the assay in routine use. If a randomized (or a balanced rotation) strip-plot design is used and any one sample within each block is present in at least two rows (or columns), then there will be sufficient data for a direct measure of any location effect. If the replicated sample is made using reference or standard material and the two replicates of the sample are prepared independently, we can monitor the variation in the sample-preparation process, although the variation associated with locations and sample preparations are now confounded (this may be appropriate if both are thought to be small but should be monitored).

The assay size is an important practical issue. An assay is defined here as the smallest group of assay data that are performed and analyzed separately. Large assays generally produce more precise estimates of relative potency, are a prerequisite for more complex models [e.g., a nonlinear mixed strip-plot model (12) requires at least two plates], and give much better estimates for any variance components estimated. Small assays produce more replicate assays with less effort. Some assay-release criteria (21) demand moderate to large numbers of assay replicates even from quite precise assays. Fortunately, at least some of the authors have recognized this as an unintended consequence of these criteria (personal communication). Large designs combined with careful statistical modeling should produce better potency estimates for less overall effort, but this approach is potentially risky given the lack of clear regulatory guidance.

The assay plate layout can change from development to validation. For production assays, we may not want or need the full curve, and we should use what we have learned from development and validation to choose the number of replicates to be used at each level of the design (i.e. how many rows/sample, how many plates/assay, etc.). Uniformity trials (4,7,10) are a simple and effective way to estimate several sources of variation and check for location effects. During validation, we emphasize estimating variation from several sources as it affects the estimated relative potency. The sources of variation that may be of interest include assays within analyst and day, batches of cells, analysts, equipment, days, and laboratories. It is often valuable to analyze these sources of variation in detail to better understand an assay system, and then summarize these as repeatability (variation among replicate samples close together in time by a single analyst using a single set of equipment), intermediate precision (variation within a laboratory across days or analysts), and reproducibility (variation among laboratories) (16). Factors that are expected to be important are deliberately varied as part of a robustness study where we generally use small ranges on the independent variables and expect little effect on the measured potency as described in Torbeck (17).

The way a layout is used may change from validation to production assay. Only changes that the validation experiment can support with valid statistical inference will be permissible between validation and production use. Assuming there are replicate assays within each analyst and day, we can estimate the repeatability directly. Similarly, if there are replicates of a sample within each assay, we can estimate within assay variation in potency directly. These estimates of variation allow us to reliably predict the performance of the assay system with various numbers of replicates at each level where we have direct replication. For example, with two replicate samples at potency 1.0 in each assay, we can predict the precision of potency when one, two, three, or more within assay replicates are combined. Similarly, if there are replicate assays within analyst and day, we can use the variation among these replicates to predict the precision of potency with one, two, three, or more replicate assays during production use.

## DESIGN OF THE VALIDATION EXPERIMENTS

### Precision and Accuracy Experiment

The general scheme is to construct artificial samples at each of the several levels of true potency using reference materials (so the true potency is known), and then assay these artificial samples repeatedly to assess the accuracy and precision of the assay system across a relevant range of potencies. For an assay intended to support a product specification of 80% to 125%, we might use true potencies of 0.64, 0.80, 1.0, and 1.25. On each of at least four days (8–20 days would be much better), at least two analysts assay each "sample" in two separate assays. If an assay can contain four or more samples, the validation experiment does not require additional assays to study multiple true potencies. The mean of the log potencies are then compared with the nominal log potencies to determine the accuracy of the assay separately at each true potency studied. The bias can be reported as a percent difference on potency scale. To assess the precision of the assay system, a variance component analysis should be conducted at each level of true potency (9,10,16), to produce a summary similar to Table 1. If the variance components for the different samples (true potencies) are comparable and either they come from independent assays (in other words, each assay contains only reference and a single test sample) or a blocked analysis is used, then it is reasonable to use pooled estimates for each variance component as a precision summary instead of the "worst case" estimate in the right-hand column of Table 1. It can be very helpful to plot the estimated log potency versus the nominal log potency; this gives a "calibration curve" for log potency. We can use this calibration curve to illustrate the linearity and range of the bioassay. Another useful summary illustrates the precision that can be expected and the assay specifications that can be supported for various numbers of assays/run and runs (Table 2). In Table 2, we use only two variance

**Table 1** Variance Component Analysis Results for Within Run (Repeatability) and Between Run (Intermediate Precision) Variability for Each of the Several True Potencies

| | Potency 0.64 | Potency 0.80 | Potency 1.0 | Potency 1.25 | "Worst case" |
|---|---|---|---|---|---|
| Between run variance | 0.00326 | 0.003 | 0.001 | 0.003 | 0.00326 |
| Within run variance | 0.0004 | 0.0003 | 0.004 | 0.0002 | 0.004 |

*Note*: For balanced designs, these can be estimated at each true potency using analysis of variance method of moments estimators. The "worst case" column contains the largest value observed across the range of true potencies used for each of within and between run.

components, assays and runs (within run and between run, otherwise known as intermediate precision and repeatability). Note that the specification limits as listed in Table 2 do not allow any variation in the product itself (all lots are assumed to be at potency 1.0) and does not leave room in the specification for any departure from perfect product stability; a statistical approach that adjusts for these sources of variation is described by Dillard (22). To compute the relative standard deviation (RSD) in Table 2, we exploit the fact that potency is nearly log-normally distributed; hence, the log of potency is nearly normal in distribution,

**Table 2** Expected Total RSD as an Intermediate Precision for Various Numbers of Runs and Assays/Run Based on the Worst-Case Values in Table 1

| Runs ($i$) | Assays/run ($m$) | Total RSD | Specification supported |
|---|---|---|---|
| 1 | 4 | 6.5 | 19.50 |
| 1 | 5 | 6.38 | 19.14 |
| 2 | 2 | 4.95 | 14.86 |
| 3 | 5 | 3.64 | 10.91 |

*Note*: The specification supported column uses a simple 3*RSD limit, a tolerance interval approach may be even more appropriate.
*Abbreviation*: RSD, relative standard deviation.

$$\begin{aligned}&\mathrm{Log}(R) = \beta + \varepsilon,\\ &Y = e^{\beta+\varepsilon} = e^{\beta}e^{\varepsilon}\\ &\text{with } \varepsilon = \text{Total SD} = \sqrt{\frac{\sigma^2_{\text{run}}}{n} + \frac{\sigma^2_{\text{assay}}}{n*m}} = \sqrt{\frac{0.00326}{n} + \frac{0.004}{n*m}}\end{aligned}$$

where $n$ is the number of runs and $m$ is the number of assays/run, and the variance estimates are taken from Table 2. Because we compute the variance components on log potency, ε is in log potency units. The percent RSD, a multiplicative error on potency scale, is computed as %RSD = 100 × exp(Total SD – 1). A substantial advantage of the summary in Table 2 is that it informs the discussion between assay managers and product managers about how to set product specifications. In particular, this approach makes it very clear how much routine assay effort is needed to support various specifications. If the effort needed to support a mandated specification is judged to be excessive, then the assay can be improved to yield more precise estimates of log potency.

## Robustness Experiment

A separate experiment to study robustness can efficiently check that the allowed ranges for many of the assay inputs are appropriate. In a robustness experiment, we are seeking to demonstrate that at least several factors and interactions among these factors have no important effect on the assay. By using a minimal fractional factorial design [i.e., a Plackett–Burman design (10)], we can check on all factors and many interactions. Note that our goal in the validation of robustness is not to study interactions among the factors we are studying, but to confirm that neither the factors nor any interactions among the factors have effects that are large enough to be of concern. Because the goal here is to show that no factor or interaction is important, we must determine prior to the experiment the effect sizes that we consider unimportant. These effect size limits establish an indifference zone. In an assay system supporting product release to a specification of 80% to 125%, it may be quite reasonable to have an indifference zone from –6% to 6% due to variation in assay inputs; in doing so, it would be appropriate to demand slightly higher precision from the assay system. It would be a mistake to simply test all factors for significant effects; instead a confidence interval should be put on each factor's effect. If all factors have confidence intervals that lie entirely within the indifference zone, then the assay can be considered robust. This approach is a simple generalization of a recently proposed and substantially improved way to assess similarity in bioassay (14,15), and is quite different from the usual analysis approach in a fractional factorial (17). The critical difference here is that when we are seeking to show that factors are not important, the analysis approach using confidence intervals on effect sizes compared with an appropriately selected indifference zone is more appropriate than conventional hypothesis tests (14,15).

## SUMMARY

Biological assays are often noisy and laborious. With careful application of experimental design, cell culture bioassays can be made quite accurate and precise. The core information needed for validation can come from two experiments. One experiment studies accuracy and precision followed by a variance component analysis and a summary table that describes the expected performance of the system at various levels of replication. A second experiment uses a minimal fractional factorial design to study robustness, followed by a comparison of confidence intervals on effect sizes with a previously established indifference zone.

## REFERENCES

1. Finney DJ. Statistical Method in Biological Assay (3rd ed.). London: Charles Griffin & Co., 1978.

2. Chapter 5.3. Statistical Analysis. In European Pharmacopoeia 5.0, 2004; 475–504.

3. Design and analysis of biological assays <111>. In: USP 30 NF 25. Rockville, MD: United States Pharma Convention, 2007: 120–132.

4. Box GEP, Draper NR. Empirical Model-Building and Response Surfaces. NY: Wiley, 1987.
5. Morgan BJT. Analysis of Quantal Response Data. NY: Chapman and Hall, 1992.
6. McCullagh P, Nelder JA. Generalized Linear Models, 2nd ed. NY: Chapman and Hall, 1989.
7. Cox DR. Planning of Experiments. NY: Wiley, 1958.
8. Federer WT. Experimental Design. Calcutta: Oxford & IBH, 1955.
9. Milliken GA, Johnson DE. Analysis of Messy Data, Volume I: Designed Experiments. London: Chapman and Hall, 1992.
10. Montgomery DC. Design and Analysis of Experiments, 6th ed. NY: Wiley, 2005.
11. Lansky D. Validation of bioassays for quality control. In: Brown F, Mire-Sluis AR, eds. Biological Characterization and Assay of Cytokines and Growth Factors. Dev Biol Stand. Basel: Karger, 1999; 97:157–168.
12. Lansky D. Strip-plot designs, mixed models, and comparisons between linear and non-linear models for microtitre plate bioassays. In: Brown W, Mire-Sluis AR, eds. The Design and Analysis of Potency Assays for Biotechnology Products. Dev Biol. Basel: Karger, 2002; 107:11–23.
13. Plikaytis BD, Holder PF, Pais LB, Maslanka SE, Gheesling LL, Carlone GM. Determination of parallelism and nonparallelism in bioassay dilution curves. J Clin Microbiol 1994; 32(10): 2441–2447.
14. Callahan, Janice D, Sajjadi, Nancy C. Testing the null hypothesis for a specified difference – the right way to test for parallelism. Bioprocessing J 2003; (2):1–6.
15. Hauck WW, Capen R, Callahan JD, et al. Assessing parallelism prior to determining relative potency. PDA J Pharm Sci Technol 2005; 59(2):127–137.
16. Schofield TL. Assay validation. In: Chow SC, ed. Encyclopedia of Biopharmaceutical Statistics. NY: Marcel Dekker, 2000: 21–30.

17. Torbeck LD. Ruggedness and robustness with designed experiments. Pharm Technol 1996; 20(3):1–3.

18. Torbeck LD, Branning RC. Designed experiments – a vital role in validation. Pharm Technol 1996; 20(6):1–4.

19. Hurlbert, Stuart H. Pseudo replication and the design of ecological field experiments. Ecol Monogr 1984; 54(2):187–211.

20. Mire-Sluis AR, Gerrard T, Gaines Das R, Padilla A, Thorpe R. Biological assays: their role in the development and quality control of recombinant biological medicinal products. Biologicals 1996; 24(4):351–362.

21. Thorpe R, Wadhwa M, Page C, Mire-Sluis A. Bioassays for the characterisation and control of therapeutic cytokines; determination of potency. In: Brown F, Mire-Sluis AR, eds. Biological Characterization and Assay of Cytokines and Growth Factors. Dev Biol Stand. Basel: Karger, 1999; 97:61–71.

22. Dillard RF. Statistical approaches to specification setting with application to bioassay. In: Brown W, Mire-Sluis AR, eds. The Design and Analysis of Potency Assays for Biotechnology Products. Dev Biol. Basel: Karger, 2002; 107:117–127.

# 5

# Use of Experimental Design Techniques in the Qualification and Validation of Plasma Protein Manufacturing Processes

**SOURAV KUNDU**[†]

Technical Operations
Aventis Behring
Bradley, Illinois, U.S.A.

[†] Currently at Amgen, Inc., West Greenwich, Rhode Island, U.S.A.

## BACKGROUND

Purified proteins of high therapeutic value can be obtained from human plasma using the fractionation technology originally proposed by Dr. Edwin J. Cohn of Harvard Medical School in early 1940s (1). The process employs various combinations of alcohol concentration, pH, ionic strength, temperature, and time to separate a number of fractions from human plasma. These fractions are rich in various proteins, such as fibrinogen, coagulation factors, von Willebrand Factor, immunoglobulins, albumin, $\alpha_1$ antitrypsin, and the like. Various downstream purification steps are then applied to purify each protein of interest in its therapeutic dosage form. Some of these downstream processes also employ various virus inactivation/reduction technologies to ensure safety of the plasma-derived products. There are many variations of the fractionation process utilized by the major manufacturers—some employing different agents for precipitation, and some using various upstream adsorption and separation steps. The fractionation industry boasts its long history of therapeutic success, product reliability, and a high degree of product-safety record.

The purification processes used to obtain therapeutic plasma proteins at industrial scale are old and established. Often, they also lack the complete package of necessary process development and validation data when held against today's standards. These data voids can be successfully backfilled by dividing a complex purification process into manageable process modules; constructing qualified down-scale models of these modules; performing designed experiments that include systematically selected process parameters at their ranges; and evaluating the quality attributes of the intermediate or final product. The design of experiment techniques provide a more cost-effective and systematic way to study these parameters when compared with other forms of experimentation, as these processes are multifactorial with complex inter-relationships between process parameters (2). The down-scale studies also provide an excellent foundation for a well-designed process validation protocol and a successful validation study.

When developing a purification process for a new plasma protein drug substance, designed experiments can be highly useful in process-parameter screening, optimization and demonstration of process robustness, similar to its use in the mainstream pharmaceutical industry described in this book by other contributors. Generally, biological products are complex mixtures with high levels of heterogeneity. Well-thought out designed experiments are the only reasonable means to resolve such complex relationships and obtain useful data.

## VALIDATION STRATEGIES FOR EXISTING PLASMA FRACTIONATION PROCESSES

In 1999, a Technical Workshop was held by the Parenteral Drug Association (PDA) to come up with a reasonable strategy for the validation of existing plasma fractionation processes. The workshop and subsequent effort by many led to a set of guidelines for the validation of existing plasma fractionation processes (3) that are shown below:

- Divide the manufacturing process into a number of "Process Modules;"
- Define process steps for each process module;
- Identify/characterize process intermediate(s);
- Identify/define process control parameter(s);
- Identify/meet process validation requirements for process step(s);
- Perform process validation for steps where validation is lacking;
- Compile and prepare process validation document.

As emphasized by the task force, down-scale qualification studies to establish robustness of existing processes are viewed as integral to this validation approach. In the present article, we will describe the strategy that can be used to generate the precursor information leading to down-scale studies and the use of experimental design approach in these studies. We will provide some practical examples of cases where meaningful data were generated to compliment validation activities.

## Strategies for Down-Scale Process Qualification

The development of a qualified down-scale model of a process module is integral to the approach of process validation using bench-scale experiments, as described earlier. We have developed down-scale models of process steps ranging from various types of process chromatography for protein purification to separation by precipitation and filtration. These down-scale models have been utilized to evaluate the effects of relevant process parameters on product-quality attributes. The normal logical sequence of process development, of course, is bench scale to pilot scale to full scale. However, for many plasma protein purification processes, a reverse order needs to be followed. As licensed full-scale processes already exist, the full-scale process steps need to be scaled down to construct small process models in order to evaluate the robustness of process parameters on the product without impacting full-scale production. These models can also be utilized to evaluate process changes, improvements, and optimizations easily and economically.

The models can be scaled to various sizes to fit the needs of the experiment. For example, a very small-scale nanofiltration system, such as a Planova P-15 hollow-fiber cartridge with 0.001-m$^2$ surface area (Asahi Kasei Corporation, Japan), can be used to study virus retention capabilities of a virus reduction step in a biological manufacturing process, whereas a scaled-up version of the same system with a surface area of 0.01 m$^2$ provides an excellent way to study the nanofiltration process variables. In a nanofiltration validation study, a feed sample is typically spiked with a known quantity of a model virus. The mixture is filtered under the expected process conditions, and virus retention is measured. Due to the cost and hazards of working with model viruses, the smallest validated down-scale model developed with a 0.001-m$^2$ nanofilter is preferable. However, for demonstration of the robustness of process parameters, a model consisting of a larger 0.01-m$^2$ nanofilter is desirable as it allows the filtration of a larger quantity of material for all of the testing needed to demonstrate product-quality attributes. Both systems are directly scalable to the production scale 1.0-m$^2$ cartridge manufactured by Asahi Kasei. It should be noted that such nicely scalable systems do not exist for all applications; therefore, models often have to be crafted, and at times approximated, for other process steps, such as chromatography, depth filtration, pasteurization, and the like.

Figure 1 provides a logical step-by-step approach to process qualification using down-scale model and bench experiments. When validating an existing process module, one has to decide whether to perform only a target validation, or evaluate process ranges. We have taken the approach of qualifying process parameter ranges at bench-scale experiments, and validating the corresponding targets in full scale. Our effort for the qualification of the critical parameter ranges for existing process modules generally start with a thorough evaluation of what is known about the process step of interest by consulting various sources of information, such as manufacturing procedures, development data, license documents, interview of operators, and the like. We have successfully employed cause–effect relationships and decision-making tools, such as fish-bone diagrams, to assist us with capturing all important process information. Figure 2 provides an example of one such attempt where we identified and captured all possible process information and their sources for an affinity chromatography purification step for a purified coagulation factor production process.

Once enough process knowledge has been acquired, the next step involves the development and qualification of an appropriate down-scale model where the parameters of interest can be effectively studied. We normally qualify our down-scale models by processing a small portion of the input

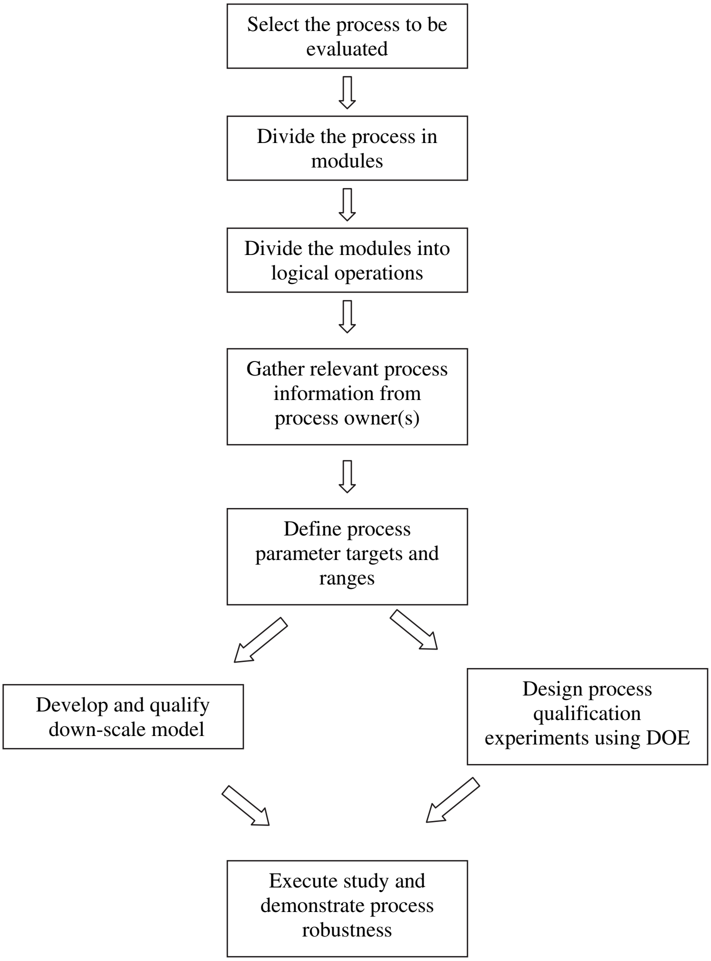

Select the process to be evaluated
Divide the process in modules
Divide the modules into logical operations
Gather relevant process information from process owner(s)
Define process parameter targets and ranges
Develop and qualify down-scale model
Design process qualification experiments using DOE
Execute study and demonstrate process robustness

**Figure 1** Process qualification strategy using bench-scale process models.

intermediate obtained from manufacturing through the down-scale model, performing appropriate assays on the output, and comparing the results with those obtained from actual product processed by manufacturing at the corresponding step. After an appropriate down-scale model has been developed and qualified, preliminary screening studies can begin. In this phase of experimentation, the goal is to evaluate the

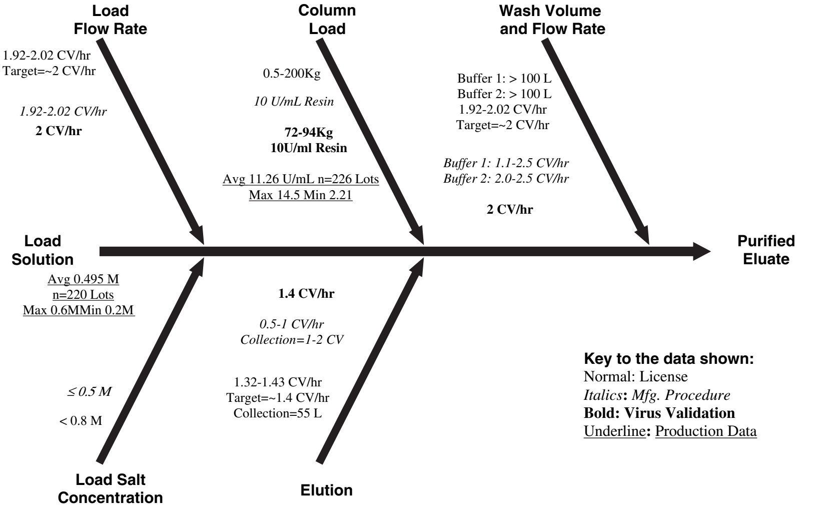


**Figure 2** Cause-and-effect (fish-bone) diagram for affinity chromatography purification.

effects of a wide array of process parameters with fairly wide ranges by using economical forms of experimental design and less than full spectrum of testing in order to keep the time and resource requirements within limits. Using a number of experimental data analysis tools, it is possible to rank the key parameters in the order of their importance on how they influence the quality of the process output. An example of this approach is described later in this article. Once the key parameters have been identified, we evaluate whether the ranges tested are appropriate for inclusion in a subsequent comprehensive process robustness study. For example, if we have evaluated the ranges for some key parameters that are well outside the normal manufacturing practice and, at these extremes, the quality of the output is outside our acceptable limits, then we would normally consider narrowing the ranges in subsequent robustness evaluations. Alternatively, if we were too conservative during our screening experiments in the selection of the range of a process parameter, and the effect of this process parameter on the quality of the process output is minimal, we would expand the range of this parameter in our subsequent robustness study. It is sometimes advisable to conduct additional screening experiments to gain enough confidence that, with the modifications in parameter ranges, there will be no surprises when more comprehensive robustness studies are performed.

We conduct comprehensive robustness studies under rigorous protocols. Acceptance criteria are decided and approved up-front. All process discrepancies are recorded and evaluated against the objective of the study. Resource availability permitting, we often employ factorial experimental design with enough power that permits the evaluation of all main effects and most interaction effects, keeping confounding to a minimum. The results of the experiments are then compared against the acceptance criteria. If all acceptance criteria are met, the process is declared qualified and robust within the ranges of the parameters tested.

## Use of Down-Scale Process Qualification Approach for Establishing Process Robustness

### Example 1: Robustness of an Affinity Chromatography Step for Purification of a Coagulation Factor

Our manufacturing process for the purification of a protein intermediate employs an affinity chromatography step where the protein is captured using a monoclonal antibody coupled to a sepharose resin. The protein is then eluted using an elution agent. In order to establish robustness of this step, we first developed and qualified a down-scale model using a 90-mL size reproduction of the industrial scale affinity column. The geometry of the column and the conditions were selected such that the linear flow rate could remain consistent between the down-scale and the full-scale model. We then listed all the parameters that are relevant at this step along with their ranges. We utilized various information sources during this effort, including manufacturing procedures (current and past versions), license documents, various development and scientific reports, actual production data, and verbal information exchange with production personnel. The cause-and-effect (fish bone) diagram (Fig. 2) discussed earlier compiled the gathered information.

The next step was to select the parameters and ranges we wanted to study in our down-scale model. Here, a careful scientific judgment had to be made as we could study only so many parameters in a reasonable amount of time. We designed an initial screening study using a $2^{(4\text{-}1)}$ fractional factorial model. We decided to study the effects of four process parameters: load amount, load flow rate, load salt concentration, and wash flow rate at their selected high and low settings as shown in Figure 3. The high and low settings of the process parameters were chosen based on the license documents, manufacturing process, and actual production data (including various deviations) to accommodate normal process fluctuations. The model was constructed using a statistical software package JMP (4). Table 1 shows the design table along with specific high and low settings of the selected process parameters. Two output response variables, % yield and specific activity of the protein, were measured. The experimental results were analyzed using JMP. The results indicated that by loading more protein on the column, the yield at the step could be increased. The yield was also related to the salt concentration of the load solution and a higher salt concentration resulted in a slightly lower yield. The effect plots for this analysis are shown in Figure 4. The analysis of the results was also able to reveal an interaction effect between the load amount and the load salt concentration (interaction plots not shown). By selecting a fractional factorial design for this initial screening study, we were able to economize on the time and resource requirements for the experiments.

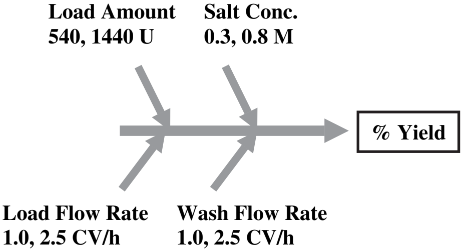


**Figure 3** Process conditions studied for immunoaffinity chromatography design of experiments (DOE) study.

**Table 1** Design Table for Screening Studies on Affinity Purification of a Protein Using Monoclonal Immunoaffinity Column ($2^{4\text{-}1}$ Fractional Factorial Design)

| | | Process parameters | | | |
|---|---|---|---|---|---|
| Run | Pattern | Load amount (U) | Salt conc. (M) | Load flow rate (CV/hr) | Wash flow rate (CV/hr) |
| 1 | −−−− | 540 | 0.3 | 1.0 | 1.0 |
| 2 | +−−+ | 1440 | 0.3 | 1.0 | 2.5 |
| 3 | −+−+ | 540 | 0.8 | 1.0 | 2.5 |
| 4 | ++−− | 1440 | 0.8 | 1.0 | 1.0 |
| 5 | −−++ | 540 | 0.3 | 2.5 | 2.5 |
| 6 | +−+− | 1440 | 0.3 | 2.5 | 1.0 |
| 7 | −++− | 540 | 0.8 | 2.5 | 1.0 |
| 8 | ++++ | 1440 | 0.8 | 2.5 | 2.5 |

Pattern column indicates the high (+) and the low (−) settings of the selected process parameters.

However, we sacrificed on the ability of this design to resolve all interaction effects. In fact, in the $2^{(4\text{-}1)}$ design, the main effects are confounded with the three factor interactions, and two factors interactions are confounded with the other two factor interactions. Ordinarily, this does not result in any serious problem since, for most processes, three factor interactions are considered small and are generally neglected. Thus, the interaction between the load amount and the load salt concentration

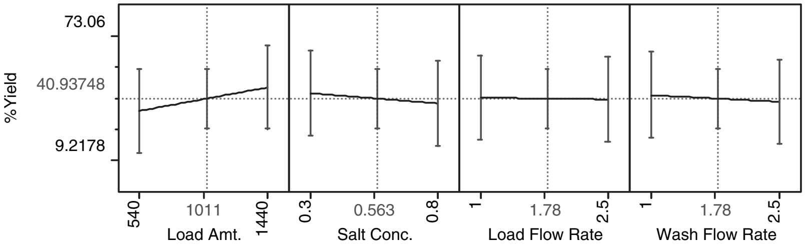


**Figure 4** Effects plots for immunoaffinity chromatography screening study.

is confounded with the interaction between load flow rate and wash flow rate. However, as the load amount and the load salt concentration are themselves larger, it is logical to assume that the interaction is between load amount and load salt concentration and not load flow rate and wash flow rate. Therefore, although from the results of this study we could not conclusively make a judgment on the exact nature of the interactions, it was reasonable to surmise that interaction effects are involved in the process performance.

The qualification study for the affinity purification step was carried out using the same down-scale process model used for the screening studies. For the qualification study, the experimental design was expanded to a full factorial model using four process variables studied earlier. The use of the full factorial model doubled the number of experimental runs but provided the statistical power to resolve all the main and interaction effects. Table 2 shows the design table along with the high and low settings of the selected process parameters. Adjustments in the settings for the load amount were made to accommodate the possibility to load more protein on the column in order to increase the batch size, and the settings for load salt concentration were also refined based on the findings from the screening study to minimize the likelihood of any influence of this parameter on the results. The results of the qualification study were analyzed by the statistical software JMP. The results confirmed the robustness of the process in terms of yield, as this parameter stayed virtually unchanged with changes in chromatography flow rate variables (Fig. 5). There was a decrease in the yield when higher load amount was used. The decrease in yield was due to exceeding of the capacity of the column, supported by the fact that active coagulation factor could be detected in the column flow-through fractions at high load conditions. The specific activity (Fig. 6) was higher with higher load, perhaps due to the fact that the affinity matrix preferentially bound coagulation factor molecules in their native form with higher activity. The activation products remained low for all the conditions studied. The statistical analysis also confirmed the presence of an interaction between the column load amount and salt concentration in

**Table 2** Design Table for Process Qualification Study for Affinity Purification of a Protein Using Monoclonal Immunoaffinity Column ($2^4$ Full Factorial Design with Center Points)

| | | Process parameters | | | |
|---|---|---|---|---|---|
| Run | Pattern | Wash flow rate (CV/hr) | Load flow rate (CV/hr) | Salt conc. (moles) | Load amount (U) |
| 1 | − − − − | 1.0 | 1.0 | 0.4 | 540 |
| 2 | + − − − | 2.5 | 1.0 | 0.4 | 540 |
| 3 | − + − − | 1.0 | 2.5 | 0.4 | 540 |
| 4 | + + − − | 2.5 | 2.5 | 0.4 | 540 |
| 5 | − − + − | 1.0 | 1.0 | 0.8 | 540 |
| 6 | + − + − | 2.5 | 1.0 | 0.8 | 540 |
| 7 | − + + − | 1.0 | 2.5 | 0.8 | 540 |
| 8 | + + + − | 2.5 | 2.5 | 0.8 | 540 |
| 9 | − − − + | 1.0 | 1.0 | 0.4 | 1738 |
| 10 | + − − + | 2.5 | 1.0 | 0.4 | 1738 |
| 11 | − + − + | 1.0 | 2.5 | 0.4 | 1738 |
| 12 | + + − + | 2.5 | 2.5 | 0.4 | 1738 |
| 13 | − − + + | 1.0 | 1.0 | 0.8 | 1738 |
| 14 | + − + + | 2.5 | 1.0 | 0.8 | 1738 |
| 15 | − + + + | 1.0 | 2.5 | 0.8 | 1738 |
| 16 | + + + + | 2.5 | 2.5 | 0.8 | 1738 |
| 17 | 0 0 0 0 | 1.75 | 1.75 | 0.6 | 1139 |
| 18 | 0 0 0 0 | 1.75 | 1.75 | 0.6 | 1139 |
| 19 | 0 0 0 0 | 1.75 | 1.75 | 0.6 | 1139 |
| 20 | 0 0 0 0 | 1.75 | 1.75 | 0.6 | 1139 |

*Note*: Pattern column indicates the high (+) and the low (–) settings of the selected process parameters. The pattern 0000 indicates that the parameter settings are at the center point of the ranges.

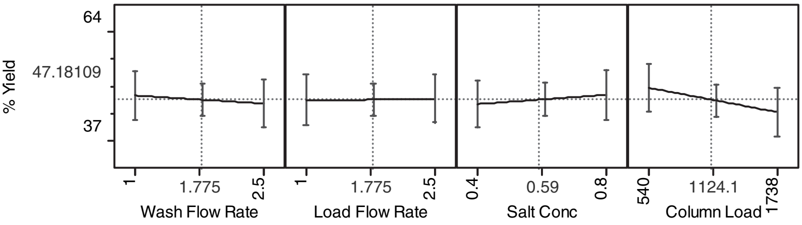


**Figure 5** Effects plot (% yield) for immunoaffinity chromatography process qualification study.

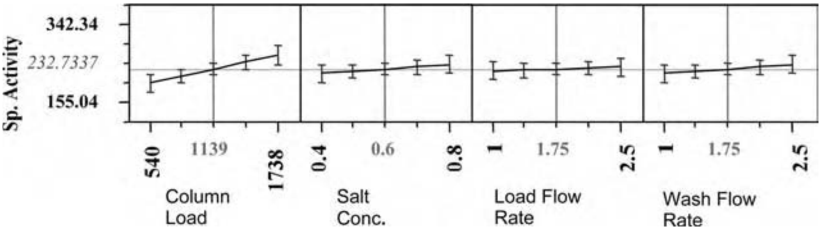
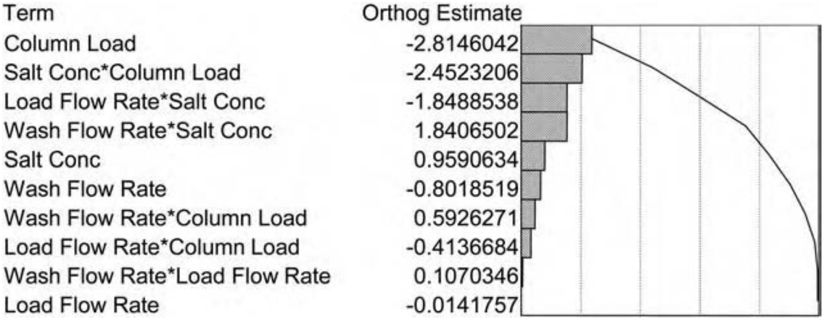


**Figure 6** Effects plot (specific activity) for immunoaffinity chromatography process qualification study.

the load demonstrated by a Pareto plot showing the relative importance of main effects and interaction effects (Fig. 7). The yield decrease was more substantial when the load amount was increased at higher salt concentration versus at lower salt concentration. The primary conclusions we made from these studies are the following:

1. The affinity chromatography process can be considered robust within the process parameter ranges studied using a qualified down-scale model.
2. The factors that had influence on yield were load amount and load salt concentration. With respect to column load amount, care should be exercised to ensure that the capacity of the column is not exceeded.



| Term | Orthog Estimate |
|---|---|
| Column Load | -2.8146042 |
| Salt Conc*Column Load | -2.4523206 |
| Load Flow Rate*Salt Conc | -1.8488538 |
| Wash Flow Rate*Salt Conc | 1.8406502 |
| Salt Conc | 0.9590634 |
| Wash Flow Rate | -0.8018519 |
| Wash Flow Rate*Column Load | 0.5926271 |
| Load Flow Rate*Column Load | -0.4136684 |
| Wash Flow Rate*Load Flow Rate | 0.1070346 |
| Load Flow Rate | -0.0141757 |

**Figure 7** Pareto plot demonstrating the relative importance of main effects and interaction effects for immunoaffinity chromatography step.

> When loading higher amounts of load material on the column, a lower salt concentration should be maintained to minimize any potential yield loss.

### Example 2: Robustness of Pasteurization and Precipitation Steps in a Fractionation/Purification Process

The production process for protein prepared by fractionation/ purification includes a heat inactivation step to ensure viral safety of the product. The heat inactivation step is followed by a precipitation step that further purifies the product. The purpose of our study was to determine whether the process steps are robust within the parameter ranges routinely used in manufacturing.

As before, the effort involved the development of a downscale model of the steps that can be reproduced on the bench. As the precipitation step follows the heat inactivation step and may remove any aggregates formed during hearing, the two steps are clearly coupled. For example, under conditions outside manufacturing limits, aggregates may be generated in higher-than-usual quantities such that the subsequent precipitation step may not be able to effectively remove them. Instead of evaluating them separately, we therefore opted to study the two consecutive steps together as a process segment in our qualification studies in order to fully understand the interaction effects between the two steps. This approach, however, increased the number of factors that we needed to evaluate and complicated the design of experiments for the qualification study. Thus, the preliminary screening studies were first performed separately for the two steps to simplify the experimental methods.

The screening study conducted for the heat inactivation step included a full factorial experiment on the following process variables: (*i*) stabilizer concentrations; (*ii*) protein concentration during pasteurization; and (*iii*) pasteurization time and temperature combinations. The ranges of various parameters investigated are shown in Figure 8 as a cause-and-effect diagram. The amount of aggregates generated during heating was the measured effect. The study design is shown in Table 3.

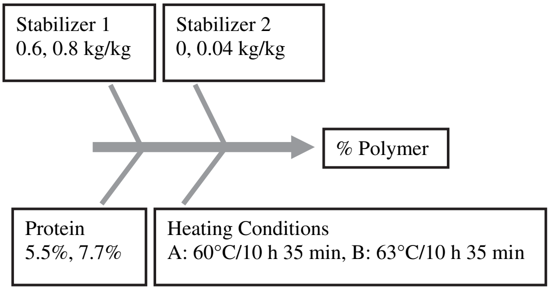


**Figure 8** Heat inactivation conditions for the screening study.

The results were analyzed using the statistical analysis software JMP. Figure 9 shows the effect plots for the parameters evaluated. As expected, the heating time and temperature combinations had the most significant influence on the

**Table 3** Design Table for the Heat Inactivation Screening Study ($2^4$ Full Factorial Experimental Design)

| | | Process parameters | | | |
|---|---|---|---|---|---|
| Run | Pattern | Protein conc. (%) | Stabilizer 1 conc. (kg/kg) | Stabilizer 2 conc. (kg/kg) | Time/ temperature |
| 1 | −−−− | 5.5 | 0.6 | 0 | A |
| 2 | +−−− | 7.7 | 0.6 | 0 | A |
| 3 | −+−− | 5.5 | 0.8 | 0 | A |
| 4 | ++−− | 7.7 | 0.8 | 0 | A |
| 5 | −−+− | 5.5 | 0.6 | 0.04 | A |
| 6 | +−+− | 7.7 | 0.6 | 0.04 | A |
| 7 | −++− | 5.5 | 0.8 | 0.04 | A |
| 8 | +++− | 7.7 | 0.8 | 0.04 | A |
| 9 | −−−+ | 5.5 | 0.6 | 0 | B |
| 10 | +−−+ | 7.7 | 0.6 | 0 | B |
| 11 | −+−+ | 5.5 | 0.8 | 0 | B |
| 12 | ++−+ | 7.7 | 0.8 | 0 | B |
| 13 | −−++ | 5.5 | 0.6 | 0.04 | B |
| 14 | +−++ | 7.7 | 0.6 | 0.04 | B |
| 15 | −+++ | 5.5 | 0.8 | 0.04 | B |
| 16 | ++++ | 7.7 | 0.8 | 0.04 | B |

Time/temperature A: 60°C/10 hr 35 min.
Time/temperature B: 63°C/10 hr 35 min.

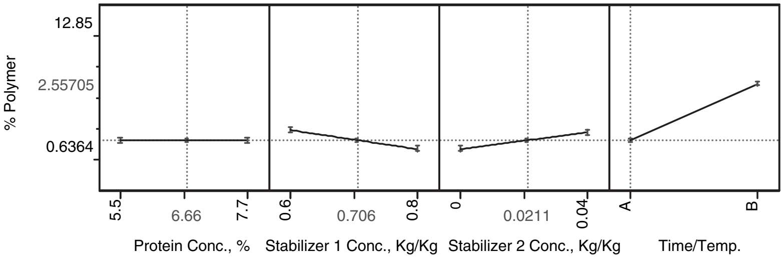


**Figure 9** Effects plots for the heat-inactivation screening study.

amount of polymer generated, followed by the concentrations of stabilizer 1 and stabilizer 2. A Pareto analysis (not shown) confirmed these findings.

The precipitation step in the purification process involves a number of process variables, including precipitant concentration, protein concentration, ionic strength, time of precipitation, temperature of precipitation, and pH during precipitation. Due to the large number of process variables, a judgment had to be made on the most important variables to include in the study. We rationalized that if we study the most important parameters in depth and hold the other parameters at their worst case or the most extreme settings, we should be able to investigate successfully the overall robustness of the process. The screening study for the precipitation step was a full factorial experiment to evaluate the following process variables: (*i*) precipitant concentration; (*ii*) pH during precipitation; and (*iii*) temperature during precipitation. The cause-and-effect diagram shown in Figure 10 describes the variables and their ranges included in the screening study. The experiments were conducted according to the design table provided by the statistical software JMP shown in Table 4. The results were analyzed and the effect plots were constructed as shown in Figure 11. All three process variables influenced percent aggregate (% polymer on the figures) in the filtrate. A Pareto analysis (not shown) determined that the pH and precipitant concentration had the highest impact on percent polymer, followed by relatively minor influences of the two-factor interaction effects and

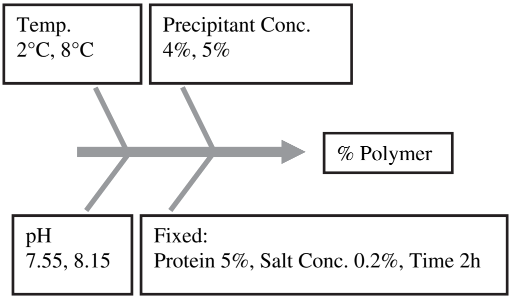


**Figure 10** Cause-and-effect diagram of process variables for precipitation screening study.

precipitation temperature. The results also indicated that higher pH and higher concentrations of precipitant at lower precipitation temperature would result in higher aggregate removal from the protein solution.

The information obtained from the screening studies provided us with the direction to take for a larger qualification study. It was clear from a scientific standpoint that heat inactivation of a protein solution generates some protein aggregates. The subsequent precipitation step purifies the

**Table 4** Design Table for Protein Precipitation Screening Study ($2^3$ Full Factorial Experimental Design with Center Point)

| | Process parameters | | | |
|---|---|---|---|---|
| Run | Pattern | Precipitation temp. (°C) | pH | Precipitant conc. (%) |
| 1 | − − − | 2 | 7.55 | 4 |
| 2 | + − − | 8 | 7.55 | 4 |
| 3 | − + − | 2 | 8.15 | 4 |
| 4 | + + − | 8 | 8.15 | 4 |
| 5 | − − + | 2 | 7.55 | 5 |
| 6 | + − + | 8 | 7.55 | 5 |
| 7 | − + + | 2 | 8.15 | 5 |
| 8 | + + + | 8 | 8.15 | 5 |
| 9 | 000 | 5 | 7.85 | 4.5 |

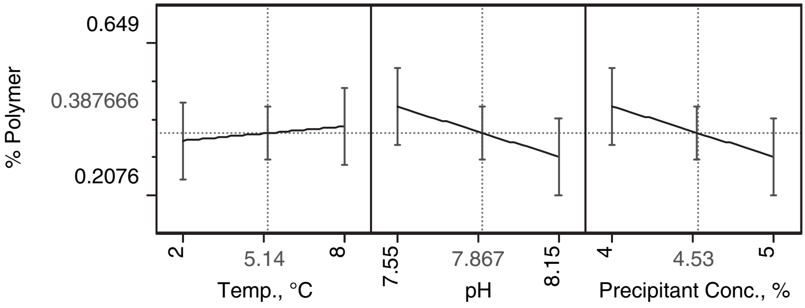


**Figure 11** Effects plots for the precipitation screening study.

protein solution. As the second step is complimentary to the first step, it was obvious that we needed to study these two steps coupled together as a process module in order to fully assess the robustness of the process. However, coupling the two steps together made the list of factors to study very long, and the next challenge was to find an experimental design that would allow us to study these factors adequately in a reasonable amount of time. We chose an L18 Hunter orthogonal array model, available in JMP software package, which allowed the evaluation of 11 factors at one time in 18 experimental runs with three factors out of 11 studied at three levels. The experimental design is shown on Table 5. For simplicity, pasteurization time and temperature were combined into one variable and studied at three discrete time/temperature conditions. We included starting intermediate lot as one of the factors, and three separate lots were included in the study. Some nonlinearity was suspected with respect to pH during precipitation, and this factor was studied at three levels. The two-level factors studied were stabilizers 1 and 2 concentrations, protein concentration during heating, protein concentration during precipitation, precipitant concentration, ionic strength during precipitation, precipitation time, and precipitation temperature. Down-scale experiments were performed with the selected factors set at the appropriate levels. The product quality attributes were measured after processing the protein to the final bulk stage and subjecting the material to

**Table 5** L18 Hunter Design Table for the Qualification Study for the Coupled Heat-Inactivation/Precipitation Steps

| | Control parameters | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Protein conc. pptn. (%) | Temp. pptn. (°C) | Pptn. time (hr) | Pptn. conc. (%) | Salt conc. pptn. (%) | Prot. conc. heat (%) | Stab. 1 conc. (kg/kg) | Stab. 2 conc. (kg/kg) | Start lot | Heat cond. | Pptn. pH |
| 1 | 4.0 | 2.0 | 2 | 5.5 | 0.25 | 7.7 | 0.04 | 0.6 | 1 | 60°C, 10 hr | 7.55 |
| 2 | 6.0 | 8.0 | 20 | 4.0 | 0.0 | 5.5 | 0.0 | 0.8 | 1 | 60°C, 10 hr | 7.55 |
| 3 | 4.0 | 8.0 | 20 | 4.0 | 0.0 | 7.7 | 0.04 | 0.6 | 1 | 60°C, 22 hr | 8.35 |
| 4 | 6.0 | 2.0 | 2 | 5.5 | 0.25 | 5.5 | 0.0 | 0.8 | 1 | 60°C, 22 hr | 8.35 |
| 5 | 4.0 | 8.0 | 2 | 4.0 | 0.25 | 5.5 | 0.04 | 0.8 | 1 | 63°C, 10 hr | 7.95 |
| 6 | 6.0 | 2.0 | 20 | 5.5 | 0.0 | 7.7 | 0.0 | 0.6 | 1 | 63°C, 10 hr | 7.95 |
| 7 | 4.0 | 2.0 | 20 | 5.5 | 0.0 | 5.5 | 0.04 | 0.8 | 2 | 60°C, 10 hr | 8.35 |
| 8 | 6.0 | 8.0 | 2 | 4.0 | 0.25 | 7.7 | 0.0 | 0.6 | 2 | 60°C, 10 hr | 8.35 |
| 9 | 4.0 | 2.0 | 20 | 4.0 | 0.25 | 7.7 | 0.0 | 0.8 | 2 | 60°C, 22 hr | 7.95 |
| 10 | 6.0 | 8.0 | 2 | 5.5 | 0.0 | 5.5 | 0.04 | 0.6 | 2 | 60°C, 22 hr | 7.95 |
| 11 | 4.0 | 8.0 | 2 | 5.5 | 0.0 | 7.7 | 0.0 | 0.8 | 2 | 63°C, 10 hr | 7.55 |
| 12 | 6.0 | 2.0 | 20 | 4.0 | 0.25 | 5.5 | 0.04 | 0.6 | 2 | 63°C, 10 hr | 7.55 |
| 13 | 4.0 | 8.0 | 20 | 5.5 | 0.25 | 5.5 | 0.0 | 0.6 | 3 | 60°C, 10 hr | 7.95 |
| 14 | 6.0 | 2.0 | 2 | 4.0 | 0.0 | 7.7 | 0.04 | 0.8 | 3 | 60°C, 10 hr | 7.95 |
| 15 | 4.0 | 2.0 | 2 | 4.0 | 0.0 | 5.5 | 0.0 | 0.6 | 3 | 60°C, 22 hr | 7.55 |
| 16 | 6.0 | 8.0 | 20 | 5.5 | 0.25 | 7.7 | 0.04 | 0.8 | 3 | 60°C, 22 hr | 7.55 |
| 17 | 4.0 | 2.0 | 2 | 4.0 | 0.0 | 5.5 | 0.0 | 0.6 | 3 | 63°C, 10 hr | 8.35 |
| 18 | 6.0 | 8.0 | 20 | 5.5 | 0.25 | 7.7 | 0.04 | 0.8 | 3 | 63°C, 10 hr | 8.35 |

a subset of the validated product release assays. The acceptance criteria were set based on the final product specifications, accounting for the variability that may enter into the picture due to the scale of operation. The protein solution prepared according to the experimental design overwhelmingly passed the acceptance criteria except for a few instances. These results were investigated further, and in all cases, assay discrepancies were noted. We could then make the conclusion that the process parameters of the heat inactivation and precipitation steps, within the ranges studied, provided material with acceptable quality, thus confirming the robustness of the process.

## Lessons Learned from the Down-Scale Qualification Studies

The down-scale process models provide an easy way to evaluate the effect of process parameters on the quality attributes of a product. These models are especially useful in a situation where a validation databank must be backfilled for existing full-scale manufacturing processes that were not validated during implementation according to the standards of today.

We presented here examples of two separate purification processes, where experimentation using qualified down-scale models and sound experimental design techniques provided important information regarding the effects of various process parameters on selected quality attributes of the output (product). In the first example describing a chromatographic purification process for a protein, the down-scale studies showed a lack of effect of the load and wash buffer flow rates on yield. In comparison, the load amount had a distinct relationship to yield. Purity, as judged by SDS-PAGE, Western blots, and specific activity remained unaffected by variation of any of the parameters within the ranges studied. Although, this lack of association seemed uninteresting to our development scientists, our validation specialists were particularly happy, as we, in effect, proved that the quality attributes of the output remain unchanged when the control parameters in this segment of the process changes. We did detect, however, several interaction effects that were of high interest to our manufacturing colleagues.

In the second example, we demonstrated that a process module might be comprised of one or more unit operations. In this example, the process module investigated comprised of several unit operations—a heat inactivation step, a precipitation step, and a filtration step. Depending on the complexity of the steps, it may be beneficial to separate them initially and evaluate their individual characteristics. However, as the steps may be coupled, as was the case in the example provided, a qualification study should evaluate the process module as a whole. In the example provided, we were able to clearly demonstrate the robustness of the module by comparing the output of the module against a subset of the established finished-product specifications.

In conclusion, the down-scale studies provided a thorough understanding of the effect of process parameters within their respective manufacturing or license ranges. The down-scale studies demonstrated the robustness of these parameters within their ranges and provided the basis for conducting full-scale validation studies. The approach for establishing process robustness by down-scale studies followed by full-scale validation appears to be a reasonable strategy for validating the existing plasma-derivative manufacturing processes.

## REFERENCES

1. Cohn EJ. The properties and functions of the plasma proteins with a consideration of the methods for their separation and purification. Chem Rev 1941; 28:395.
2. Czitrom V. One factor at a time versus designed experiments. Am Statistician 1999; 53:126.
3. Technical Workshop: Validating Plasma Fractionation Processes, PDA, February 8–9, Bethesda, MD.
4. JMP Statistical Discovery Software v5.0, SAS Institute, Cary, NC.

# 6

# Response Surface Methodology for Validation of Oral Dosage Forms

**THOMAS D. MURPHY**
T. D. Murphy Statistical Consulting, LLC
Morristown, New Jersey, U.S.A.

## INTRODUCTION

Response Surface Methodology (RSM) is a well-known statistical technique (1–3) used to define the relationships of one or more process output variables (responses) to one or more process input variables (factors) when the mechanism underlying the process is either not well understood or is too complicated to allow an exact predictive model to be formulated from theory. This is a necessity in process validation, where limits must be set on the input variables of a process to assure that the product will meet predetermined specifications and quality characteristics. Response data are collected from the process under designed operating conditions, or specified settings of one or more factors, and an empirical mathematical function (model) is fitted to the data to define the relationships between process inputs and outputs. This empirical model is then used to predict the optimum ranges of the response variables and to determine the set of operating conditions which will attain that optimum. Several examples listed in Table 1 exhibit the applications of RSM to processes, factors, and responses in process validation situations.

## EMPIRICAL MODELS IN RESPONSE SURFACE METHODOLOGY

The most useful empirical models are the second-order polynomial models and occasionally a first- or third-order polynomial. The measurement scales of the responses or factors may be transformed to another metric, such as a logarithmic scale, or

**Table 1** Applications of Response Surface Methodology

| Processes | Factors | Responses |
|---|---|---|
| Milling | Types of blades | Particle size |
| (wet granulation) | Rotation speed | Fines, oversize |
| | Milling time | Agglomeration |
| | Milling temperature | Flowability |
| | Solvent feed rate | |
| | Solvent type | |
| Blending | Blending time | Homogeneity |
| | Mixer type | |
| | Feed rate | |
| Tablet coating | Batch size | Dissolution |
| | Spray rate | Content uniformity |
| | Inlet air temperature | Appearance |
| | Ingredient levels | Coating efficiency |
| Compression | Feed rate | Weight |
| | Piston travel | Thickness |
| | Type of press | Hardness |
| | Powder feed system | Content uniformity |

a nonlinear model might be used (2,3), but these are more complicated situations and will not be discussed in this chapter.

Second-order polynomial models are versatile enough to describe most relationships between factors and responses. These models consist of an intercept term, first-order and second-order terms for each factor, and two-factor interaction terms for each combination of two factors. The second-order model is shown below for two factors and for multiple factors.

Two factors: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2$

Multiple (k) factors: $Y = \beta_0 + \sum_{i=1}^{k} \beta_i \mathrm{X}_i + \sum_{i=1}^{\mathrm{k}} \beta_{ii} X_{\mathrm{i}}^2 + \sum_{i=1}^{k\text{-}1} \sum_{j>i}^{k} \beta_{ij} X_i X_j$

## EXPERIMENTAL DESIGNS FOR RESPONSE SURFACE METHODOLOGY

Some general criteria for setting factor levels in a second-order response surface design are the following:

1. Coverage of the experimental region (factor space);
2. Symmetry about the center of the region;

3. The number of distinct factor-level combinations at least equal to the number of terms in the model;
4. An imbedded $2^{k-p}$ factorial structure is necessary for the estimation of two-factor interactions;
5. Three levels of each factor at a constant level of the other factors are required for the estimation of quadratic effects.

As shown in Table 2, the number of terms in the model grows rather quickly as the number of factors increase in the study. As the number of distinct experimental runs in the design must be equal to or greater than the number of terms on the model, a design can become costly as more factors are added to the program. For this reason, most RSM studies are usually limited to no more than five factors.

The experiment design involves selecting multiple levels of each factor to define the conditions for each experimental run, and some subject matter expert/end-user knowledge is required to select the range of the factor levels to cover. As the optimum level of the response is not known in advance, at least five factor levels should be used, and the range of these levels should be as wide as possible with the low and high levels determined by the end user's input of feasibility and safety considerations.

For two or more factors, the central composite design (CCD) is widely used and has certain advantages. The CCD

**Table 2** Number of Terms in the Second-Order Model by Number of Factors

| | Number of terms by type | | | | Total |
|---|---|---|---|---|---|
| Factors $k$ | Intercept | Linear | Quadratic | Interaction | terms |
| 1 | 1 | 1 | 1 | 0 | 3 |
| 2 | 1 | 2 | 2 | 1 | 6 |
| 3 | 1 | 3 | 3 | 3 | 10 |
| 4 | 1 | 4 | 4 | 6 | 15 |
| 5 | 1 | 5 | 5 | 10 | 21 |
| 6 | 1 | 6 | 6 | 15 | 28 |
| 7 | 1 | 7 | 7 | 21 | 36 |
| 8 | 1 | 8 | 8 | 28 | 45 |

**Table 3** Central Composite Design Conditions for Two Factors

| Run number | Factor $X_1$ | Factor $X_2$ |
|---|---|---|
| 1 | –1 | –1 |
| 2 | +1 | –1 |
| 3 | –1 | +1 |
| 4 | +1 | +1 |
| 5 | 0 | 0 |
| 6 | $-\alpha$ | 0 |
| 7 | $+\alpha$ | 0 |
| 8 | 0 | $-\alpha$ |
| 9 | 0 | $+\alpha$ |

for two factors, as shown in Table 3, is composed of a $2^2$ factorial design with four factorial run conditions (runs 1–4), one or more center points, or run conditions at the average of the low and high levels of each factor (run 5), and four star or axial points, which are run conditions where one factor is held at the center-point level and the other factor at another level, $\pm\alpha$ (runs 6–9). The factor settings are listed in "coded" form as –1 for the low factorial level, +1 for the high factorial level, and 0 for the center-point level. A graphical layout of the factor levels for these experimental runs is depicted in Figure 1.

The value of $\alpha$ is not critical, but to achieve a rotatable design for two factors, $\alpha = \sqrt{2}$, which places all the factorial

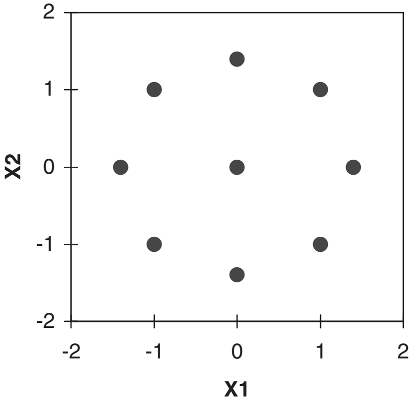


**Figure 1** Central composite design.

and star points on a circle. A rotatable design has the desirable feature that the model prediction variance is a function of the distance from the center. In general, for $k$ factors, rotatability is achieved with, $\alpha = \sqrt[4]{F}$ where $F$ is the number of factorial points in the design.

If $\alpha = 1$, the layout becomes a $3^2$ design, which is not a rotatable design, but does lend some simplicity in that only three levels of each factor are used instead of five. This may be useful if the factor levels are difficult to change, such as the installation of equipment (changing extruder screws) or the need for system equilibration when adjusting temperatures.

It is recommended that the center-point conditions be replicated to give an estimate of experimental variation. The recommended number of center points can vary, but for simplicity, it has been recommended (3) that 3–5 center points be used in a CCD. The CCD is useful for building RSM models, as the design can be conducted in a sequential order, enabling the experimenter to consider models of increasing complexity as the experimental program proceeds. For example, first run a $2^2$ factorial design with two center points, enabling the estimation of main effects and the two-factor interaction. The second phase would add the star points with two additional center points. The agreement of the two sets of center points can determine if a shift has occurred in the time between the two sets of experiments (block effect).

A non-CCD is another alternative if the program is conducted sequentially. If the results from the $2^2$ design indicate that the factor space should be extended in a particular direction, then two star points can be added to a corner as shown in Figure 2. This design also meets all the requirements for a second-order design. For four or more factors, a CCD might use an imbedded fractional factorial design plus center points and the $2k$ star points, where $k$ is the number of factors.

A CCD for three factors is listed in Table 4, and includes an imbedded $2^3$ factorial design with center points and three pairs of star points. For rotatability, $\alpha = 1.682$, as there are $F = 8$ factorial points in the design. A three-block sequential strategy for a three-factor CCD is listed in Table 5, where:

1. A $2^{3\text{-}1}$ fractional factorial with two center points gives an estimate of first-order effects;

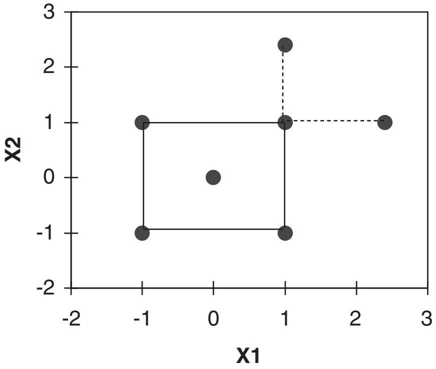


**Figure 2** Non–central composite design.

2. The other half of the $2^3$ factorial design with two center points allows for the estimation of two-factor interactions; and
3. The six star points plus two more center points allow for the estimation of the quadratic terms. The six center points give an estimate of experimental

**Table 4** Central Composite Design Conditions for Three Factors

| Run number | Factor $X_1$ | Factor $X_2$ | Factor $X_3$ |
|---|---|---|---|
| 1 | –1 | –1 | –1 |
| 2 | +1 | –1 | –1 |
| 3 | –1 | +1 | –1 |
| 4 | +1 | +1 | –1 |
| 5 | –1 | –1 | +1 |
| 6 | +1 | –1 | +1 |
| 7 | –1 | +1 | +1 |
| 8 | +1 | +1 | +1 |
| 9 | 0 | 0 | 0 |
| 10 | $-\alpha$ | 0 | 0 |
| 11 | $+\alpha$ | 0 | 0 |
| 12 | 0 | $-\alpha$ | 0 |
| 13 | 0 | $+\alpha$ | 0 |
| 14 | 0 | 0 | $-\alpha$ |
| 15 | 0 | 0 | $+\alpha$ |

$\alpha = 1.682$ for rotatability.

**Table 5** Sequential Approach for Conducting a Three-Factor Central Composite Design in Three Blocks

| $X_1$ | $X_2$ | $X_3$ |
|---|---|---|
| *Block 1* Perform a $2^{3-1}$ fraction factorial design plus 2 center points. Estimate the factor main effects (first-order effects) and the overall curvature effect. | | |
| − | − | + |
| + | − | − |
| − | + | − |
| + | + | + |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| *Block 2* Run the other half of the $2^{3-1}$ design plus 2 center points. A foldover design; estimate the three two-factor interaction effects, re-estimate main effects and overall curvature, and the block effect. | | |
| − | − | − |
| + | − | + |
| − | + | + |
| + | + | − |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| *Block 3* Add the six star points plus 2 center points, fit the full second order model, estimate block effects. | | |
| $-\alpha$ | 0 | 0 |
| $+\alpha$ | 0 | 0 |
| 0 | $-\alpha$ | 0 |
| 0 | $+\alpha$ | 0 |
| 0 | 0 | $-\alpha$ |
| 0 | 0 | $+\alpha$ |
| 0 | 0 | 0 |
| 0 | 0 | 0 |

variations and also allow the estimation of block effects.

Another important class of RSM designs, in addition to the composite designs, is the Box–Behnken design (2,3). These designs are not built up from factorial designs, but are used for second-order fitting from the start, and are more economical to use with larger numbers of factors; say, greater than four. The design layout for a three-factor Box–Behnken design is listed in Table 6.

**Table 6** The Three-Factor Box–Behnken Design

| Run | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| 1 | − | − | 0 |
| 2 | + | − | 0 |
| 3 | − | + | 0 |
| 4 | + | + | 0 |
| 5 | − | 0 | − |
| 6 | − | 0 | + |
| 7 | + | 0 | − |
| 8 | + | 0 | + |
| 9 | 0 | − | − |
| 10 | 0 | + | − |
| 11 | 0 | − | + |
| 12 | 0 | + | + |

## RESPONSE SURFACE METHODOLOGY ANALYSIS

After the experimental runs are conducted and the responses are measured, the model is fitted to the data by a regression analysis procedure in a statistical package, and two of these are listed as follows.

> MINITAB® statistical software package (Minitab, Inc., State College, Pennsylvania, U.S.A.), release for Windows® (Microsoft Corp., Redmond, Washington, U.S.A.).
>
> DESIGN EXPERT® statistical package (Stat-Ease, Inc., Minneapolis, Minnesota, U.S.A.).

Various graphs are available to aid in the interpretation of the results. The end product of an experiment design is a report listing the factor ranges necessary to produce, within statistical uncertainty, the required response levels to attain the desired product quality.

The RSM procedure will now be illustrated through three examples of increasing complexity.

### Case Study 1—Blending for Product Uniformity: A Single-Factor Response Surface Methodology

Consider a blending study to determine the optimum blending time of a mixture of solid particulates or powders. An active pharmaceutical ingredient (API) was blended with three excipients in a ribbon blender. The study was conducted by thief sampling at six different locations in the blender after a given blending time, measuring the content of the drug substance at each sampling point, and calculating the variation in API assay of the six samples as percent relative standard deviation (% RSD). The response was blend uniformity and the factor was blending time. The study was conducted for several blending times, ranging from 15 to 60 minutes. Two empirical models could be used to describe the relationship, a first-order polynomial or straight-line model and the second-order polynomial or parabolic model. The data for this study and the fitted values are listed in Table 7.

Plots of the data and fitted models are shown in Figure 3. The response "surface" in this case is a one-dimensional object—a line or a curve—and is easy to interpret graphically.

The expressions for the two candidate empirical models are:

First order: $Y = b_0 + b_1 X$
Second order: $Y = b_0 + b_1 X + b_{11} X^2$

The coefficients of the model terms are the intercept $b_0$, the slope $b_1$, and the curvature $b_{11}$. These are readily estimated by a statistical regression analysis package. The intercept

**Table 7** Blend Uniformity Study

| Blend | Uniformity (% RSD) | | |
|---|---|---|---|
| Time (min) | Observed data | 1st order fit | 2nd order fit |
| 15 | 5.8 | 5.6 | 5.7 |
| 20 | 4.5 | 4.7 | 4.5 |
| 25 | 3.6 | 3.8 | 3.6 |
| 30 | 2.7 | 2.9 | 2.9 |
| 35 | 2.2 | 2.0 | 2.4 |
| 40 | 2.2 | | 2.2 |
| 45 | 2.4 | | 2.2 |
| 50 | 2.6 | | 2.5 |
| 55 | 3.1 | | 3.0 |
| 60 | 3.5 | | 3.7 |

*Abbreviation*: RSD, relative standard deviation.

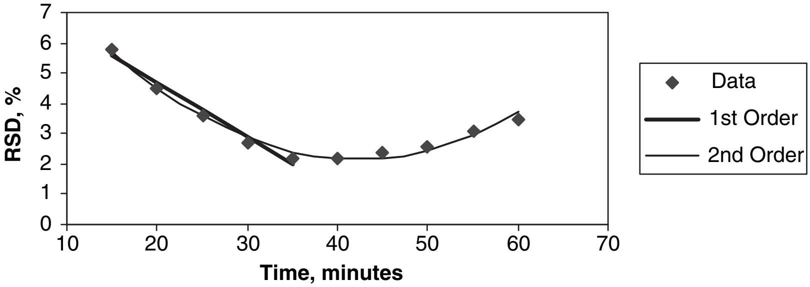


**Figure 3** Blending study.

coefficient is the predicted response value when the factor level is zero, and may not be interpretable if the factor-level range is far removed from zero. The slope coefficient is the increase in the response level attributed to a unit increase in the factor level when the relationship between $Y$ and $X$ is first order.

The curvature coefficient, although not readily interpretable numerically, can be used to estimate the factor level that gives the maximum or minimum value of the response, and this occurs at the factor level value $X = -b_1 / 2b_{11}$. (The maximum or minimum response occurs when the first derivative is equal to zero, that is, $dY/dX = b_1 + 2b_{11}X = 0$, and solving for $X$ gives the desired value.) The sign of the curvature coefficient is positive when the response goes through a response minimum and is negative for a response maximum. A handy mnemonic is when $b_{11}$ is positive, the curve is a "smile," and when negative, a "frown."

For the data in the blending study, the estimated first-order model was $Y = 8.26 - 0.18X$ over the 15- to 35-minute range. In this range, the RSD decreased by 0.18% per minute of blending time. The intercept value of 8.26% RSD predicted the RSD at zero minutes blend time, which was certainly not of interest in this type of study. In addition, the model was not at all predictive of the results past 35 minutes.

The second-order model was estimated with the output listed in Table 8, and the resulting model equation was $Y = 10.7 - 0.407X + 0.00483X^2$. The predicted minimum RSD

**Table 8** Regression Analysis Output for Blender Study[a]

| Predictor | Coef. | SE Coef. | T | P |
|---|---|---|---|---|
| *The regression equation is RSD = 10.7 – 0.407 Time + 0.00483 Time²* | | | | |
| Constant | 10.7145 | 0.3832 | 27.96 | 0.000 |
| Time | –0.40662 | 0.02237 | –18.18 | 0.000 |
| Time$^2$ | 0.0048333 | 0.0002941 | 16.43 | 0.000 |
| S = 0.1689 | R-Sq = 98.3% | R-Sq(adj) = 97.8% | | |

[a] Regression Analysis: RSD vs. Time, Time$^2$

*Abbreviations*: Coef, estimated coefficient; P, p-value; RSD, relative standard deviation; SE Coef, standard error of estimated coefficient; T, t statistic.

was attained at X = –(–0.407)/(2 × 0.00483) = 42.1 minutes. The end result of this study might state that the blending time must be maintained in the range of 37 to 47 minutes in order to obtain the maximum uniformity (minimum RSD).

## Case Study 2—Granulation Milling for Particle-Size Control: A Two-Factor Response Surface Methodology

This study dealt with a granulation milling experiment with the two factors being impeller speed and solvent addition time. The response was the product yield after screening, the desirable material passing through a 40-mesh screen and retained on a 100-mesh screen. The material passing through a 100-mesh screen, or "fines," and the material retained on a 40-mesh screen, or "overs," were discarded or reworked and represented additional processing costs at best or product loss at worst. The limiting factor levels chosen for this study were 120 to 180 rpm for the impeller speed and 65 to 85 g per minute for the solvent addition rate.

The 11-run design and resulting data are listed in Table 9. The first four runs are the imbedded $2^2$ factorial points, the next four runs are the axial or star points, and the last three runs are the center points. The factor levels are listed in the actual units ("uncoded" form) and in coded units. The value of $\alpha$ was set to 1 to minimize the number of factor levels required, instead of conducting a rotatable design

**Table 9** Granulation Milling Study Design and Results

| Factors and levels | | | | Responses | | |
|---|---|---|---|---|---|---|
| Impeller speed (rpm) | Addition rate (gpm) | Impeller speed coded | Addition rate coded | Fines % w/w | Overs % w/w | Yield % w/w |
| 120 | 65 | –1 | –1 | 0 | 12 | 88 |
| 180 | 65 | 1 | –1 | 20 | 0 | 80 |
| 120 | 85 | –1 | 1 | 26 | 14 | 60 |
| 180 | 85 | 1 | 1 | 26 | 14 | 60 |
| 120 | 75 | –1 | 0 | 13 | 15 | 72 |
| 180 | 75 | 1 | 0 | 23 | 9 | 68 |
| 150 | 65 | 0 | –1 | 10 | 4 | 86 |
| 150 | 85 | 0 | 1 | 26 | 12 | 62 |
| 150 | 75 | 0 | 0 | 18 | 10 | 72 |
| 150 | 75 | 0 | 0 | 17 | 9 | 74 |
| 150 | 75 | 0 | 0 | 19 | 11 | 70 |

where five levels of each factor would be required ($-\alpha$, –1, 0, 1, and $\alpha$ in coded form).

The three empirical models could be used as response surfaces, and from the simplest to the most complex, these were the additive (first-order) model, the interactive model, and the full-quadratic (second-order) model. The additive and interactive models could be considered as special cases of the full-quadratic model.

The additive model in two factors was $Y = b_0 + b_1X_1 + b_2X_2$ and implied that the two factors influenced the response independently and in a magnitude proportional to the two slope coefficients, $b_1$ and $b_2$. The response surface could be depicted as a plane, or as a series of straight line contours. Suppose the effects of impeller speed ($X_1$) and solvent rate ($X_2$) each increased the percent fines within the experimental range under consideration, with an estimated model $Y = -67 + 0.17X_1 + 0.8X_2$, as computed from the four factorial runs. The resulting contour plot is shown in Figure 4. The contours of constant level of percent fines are drawn over the two-dimensional factor space. The contours are shown for 10%, 15%, 20%, 25%, and 30% fines. Note that the fines steadily increase with both the impeller speed and the solvent addition rate, and the maximum

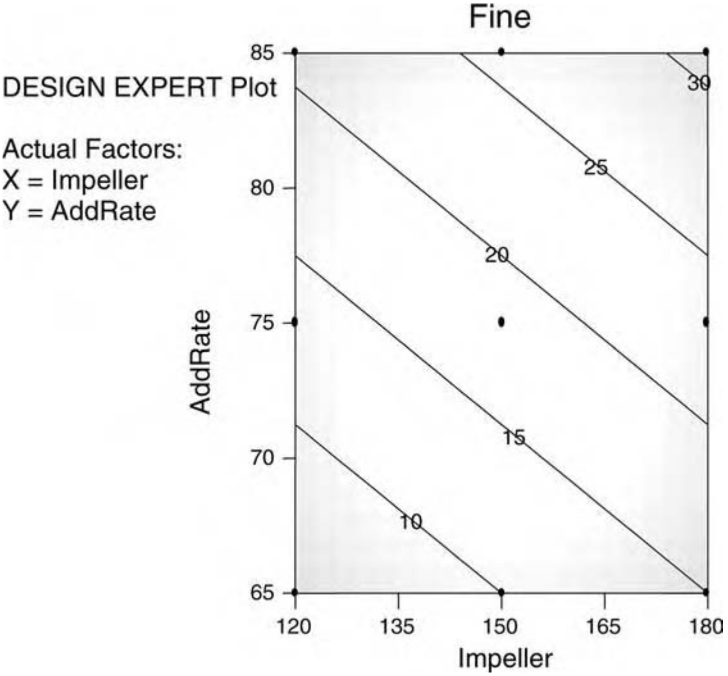


**Figure 4** An additive model contour plot for fines from the milling study.

value from this prediction occurs at the 180-rpm impeller speed and the 85-g per minute addition rate.

Another graphical aid is the three-dimensional plot shown in Figure 5, where the response surface is depicted as a grid of points representing a tilted flat plane hovering above the plane of the factor space. The contours are shown below the response surface on the plane representing the factor space. Both of these plots were generated by the Design Expert statistical package.

The interactive model in two factors is $Y = b_0 + b_1 X_1 + b_2 X_2 + b_{12} X_1 X_2$ with a fourth term that indicates that the response is also being influenced by a multiplicative combination of the two factors. If the interaction coefficient $b_{12}$ is positive, then the interaction is said to be synergistic, and the effect of the two factors on the response is greater than the

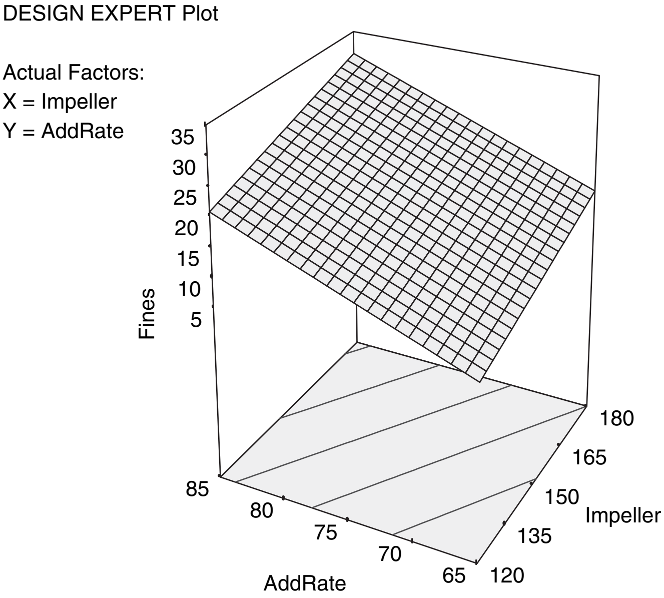


**Figure 5** An additive model three-dimensional response surface plot for fines from the milling study.

sum of the two independent linear effects of the two factors. A negative coefficient indicates a depressive multiplicative effect. The three-dimensional plot in Figure 6 corresponds to the fitted model $Y=-254.5 + 1.42X_1 + 3.3X_2 - 0.017X_1X_2$, again calculated from the results of the four factorial points. The response surface is no longer flat but now droops toward low values of the two factors and flattens toward the higher values of the two factors. The contours are a series of curves.

The full quadratic model in two factors was, $Y=b_0+b_1X_1+b_2X_2+b_{11}X_1^2+b_{12}X_1X_2+b_{22}X_2^2$, which has two added second order, or curvature terms added for each of the two factors. This model is the most versatile for response surface modeling.

The regression output is shown in Table 10A for percent yield, Table 10B for percent fines, and Table 10C for percent overs. The analysis of variance part of each table indicated that the linear and interaction terms were statistically

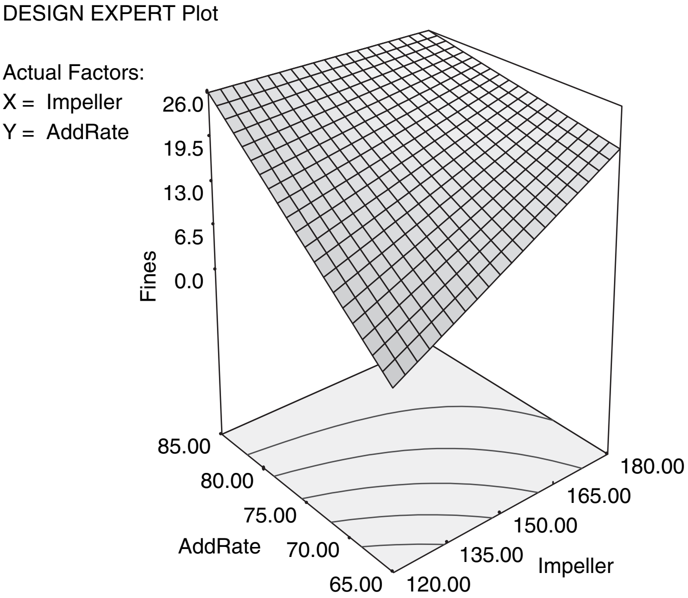


**Figure 6** An interactive model three-dimensional response surface plot for fines in the milling study.

significant for all three responses, but the quadratic terms were significant only for yield and overs. The regression model coefficients were calculated in two ways—the first based on coded factor levels and the second based on uncoded factor levels. The estimates based on coded units are known as standardized coefficients. Here the $X$ variables are standardized by first *centering* (subtracting off the mean) and then *scaling* (dividing by the range) before fitting the quadratic model. Centering reduces the correlation among the intercept, linear, and quadratic terms. Scaling allows for better comparison of the magnitude of the linear effects among factors, as all uncoded factor levels are not on the same size scale.

The contour plot for yield, shown in Figure 7, indicated that the highest yields (greater than 85%) occurred at low to mid impeller speeds, or 120 to 150 rpm, and the lowest addition rate of 65 g per minute. Although the principal response

**Table 10A** Regression Analysis Output for Milling Study Percent Yield Response: Response Surface Regression—Yield vs. Impeller, AddRate

Analysis of variance for yield

| Source | df | Seq SS | Adj SS | Adj MS | F | P |
|---|---|---|---|---|---|---|
| Regression | 5 | 920.000 | 920.000 | 184.000 | 115.00 | 0.000 |
| Linear | 2 | 888.000 | 888.000 | 444.000 | 277.50 | 0.000 |
| Square | 2 | 16.000 | 16.000 | 8.000 | 5.00 | 0.064 |
| Interaction | 1 | 16.000 | 16.000 | 16.000 | 10.00 | 0.025 |
| Residual error | 5 | 8.000 | 8.000 | 1.600 | | |
| Lack-of-Fit | 3 | 0.000 | 0.000 | 0.000 | 0.00 | 1.000 |
| Pure error | 2 | 8.000 | 8.000 | 4.000 | | |
| Total | 10 | 928.000 | | | | |

Estimated regression coefficients for yield: coded units

| Term | Coef. | SE Coef. | T | P |
|---|---|---|---|---|
| Constant | 72.00 | 0.6489 | 110.959 | 0.000 |
| Impeller | –2.00 | 0.5164 | –3.873 | 0.012 |
| AddRate | –12.00 | 0.5164 | –23.238 | 0.000 |
| Impeller*Impeller | –2.00 | 0.7947 | –2.517 | 0.053 |
| AddRate*AddRate | 2.00 | 0.7947 | 2.517 | 0.053 |
| Impeller*AddRate | 2.00 | 0.6325 | 3.162 | 0.025 |
| S = 1.265 | R-Sq = 99.1% | 1% | R-Sq (adj) = 98.3% | |

Estimated regression coefficients for yield: uncoded units

| Term | Coef. |
|---|---|
| Constant | 309.500 |
| Impeller | 0.100000 |
| AddRate | –5.20000 |
| Impeller*Impeller | –0.00222222 |
| AddRate*AddRate | 0.0200000 |
| Impeller*AddRate | 0.00666667 |

*Abbreviations*: Adj MS, adjusted mean square; Adj SS, adjusted sum of squares; Coef., estimated coefficient; df, degrees of freedom; F, f statistic; P, p-value; R-Sq, multiple coefficient of determination; R-Sq(adj), adjusted multiple coefficient of determination; S, residual standard deviation; SE Coef., standard error of estimated coefficient; Seq SS, sequential sum of squares; T, t statistic.

was yield, some additional process understanding could be obtained by evaluating the fines and overs as well, as was done in Figures 8 and 9. The predicted amount of fines (Fig. 8) was less than 5% at impeller speeds of 120 to 135 rpm and addition

**Table 10B** Regression Analysis Output for Milling Study Percent Fines Response: Response Surface Regression—Fines vs. Impeller, AddRate

Analysis of variance for fines

| Source | df | Seq SS | Adj SS | Adj MS | F | P |
|---|---|---|---|---|---|---|
| Regression | 5 | 634.000 | 634.000 | 126.800 | 317.00 | 0.000 |
| Linear | 2 | 534.000 | 534.000 | 267.000 | 667.50 | 0.000 |
| Square | 2 | 0.000 | 0.000 | 0.000 | 0.00 | 1.000 |
| Interaction | 1 | 100.000 | 100.000 | 100.000 | 250.00 | 0.000 |
| Residual error | 5 | 2.000 | 2.000 | 0.400 | | |
| Lack-of-fit | 3 | 0.000 | 0.000 | 0.000 | 0.00 | 1.000 |
| Pure error | 2 | 2.000 | 2.000 | 1.000 | | |
| Total | 10 | 636.000 | | | | |

Estimated regression coefficients for fines: coded units

| Term | Coef. | SE Coef. | T | P |
|---|---|---|---|---|
| Constant | 18.000 | 0.3244 | 55.480 | 0.000 |
| Impeller | 5.000 | 0.2582 | 19.365 | 0.000 |
| AddRate | 8.000 | 0.2582 | 30.984 | 0.000 |
| Impeller*Impeller | 0.000 | 0.3974 | 0.000 | 1.000 |
| AddRate*AddRate | –0.000 | 0.3974 | –0.000 | 1.000 |
| Impeller*AddRate | –5.000 | 0.3162 | –15.811 | 0.000 |
| S = 0.6325 | R-Sq = 99.7% | | R-Sq(adj) = 99.4% | |

Estimated regression coefficients for fines: uncoded units

| Term | Coef. |
|---|---|
| Constant | –254.500 |
| Impeller | 1.41667 |
| AddRate | 3.30000 |
| Impeller*Impeller | 4.548643E-19 |
| AddRate*AddRate | –2.26224E-18 |
| Impeller*AddRate | –0.0166667 |

*Abbreviations*: Adj MS, adjusted mean square; Adj SS, adjusted sum of squares; Coef., estimated coefficient; df, degrees of freedom; F, f statistic; P, p-value; R-Sq, multiple coefficient of determination; R-Sq(adj), adjusted multiple coefficient of determination; S, residual standard deviation; SE Coef., standard error of estimated coefficient; Seq SS, sequential sum of squares; T, t statistic.

rates of 65 to 68 g per minute. The predicted amount of overs (Fig. 9) was lowest at the higher impeller speeds, 160–180 rpm, and again low addition rates. This analysis showed a tradeoff between high production of fines at high impeller speeds and

**Table 10C** Regression Analysis Output for Milling Study Percent Overs Response: Response Surface Regression—Overs vs. Impeller, AddRate

Analysis of variance for overs

| Source | df | Seq SS | Adj SS | Adj MS | F | P |
|---|---|---|---|---|---|---|
| Regression | 5 | 202.000 | 202.000 | 40.4000 | 101.00 | 0.000 |
| Linear | 2 | 150.000 | 150.000 | 75.0000 | 187.50 | 0.000 |
| Square | 2 | 16.000 | 16.000 | 8.0000 | 20.00 | 0.004 |
| Interaction | 1 | 36.000 | 36.000 | 36.0000 | 90.00 | 0.000 |
| Residual error | 5 | 2.000 | 2.000 | 0.4000 | | |
| Lack-of-fit | 3 | –0.000 | –0.000 | –0.0000 | 0.00 | 1.000 |
| Pure error | 2 | 2.000 | 2.000 | 1.0000 | | |
| Total | 10 | 204.000 | | | | |

Estimated regression coefficients for overs: coded units

| Term | Coef. | SE Coef. | T | P |
|---|---|---|---|---|
| Constant | 10.000 | 0.3244 | 30.822 | 0.000 |
| Impeller | –3.000 | 0.2582 | –11.619 | 0.000 |
| AddRate | 4.000 | 0.2582 | 15.492 | 0.000 |
| Impeller*Impeller | 2.000 | 0.3974 | 5.033 | 0.004 |
| AddRate*AddRate | –2.000 | 0.3974 | –5.033 | 0.004 |
| Impeller*AddRate | 3.000 | 0.3162 | 9.487 | 0.000 |
| S = 0.6325 | R-Sq = 99.0% | | R-Sq(adj) = 98.0% | |

Estimated regression coefficients for overs: uncoded units

| Term | Coef. |
|---|---|
| Constant | 45.0000 |
| Impeller | –1.51667 |
| AddRate | 1.90000 |
| Impeller*Impeller | 0.00222222 |
| AddRate*AddRate | –0.0200000 |
| Impeller*AddRate | 0.0100000 |

*Abbreviations*: Adj MS, adjusted mean square; Adj SS, adjusted sum of squares; Coef., estimated coefficient; df, degrees of freedom; F, f statistic; P, p-value; R-Sq, multiple coefficient of determination; R-Sq(adj), adjusted multiple coefficient of determination; S, residual standard deviation; SE Coef., standard error of estimated coefficient; Seq SS, sequential sum of squares; T, t statistic.

high production of overs at low impeller speeds, but the increase in fines appeared to be the dominating influence on yield.

The overall conclusion is to keep low addition rates, possibly investigating lower addition rates, with low to medium impeller speeds in order to obtain maximum yields.

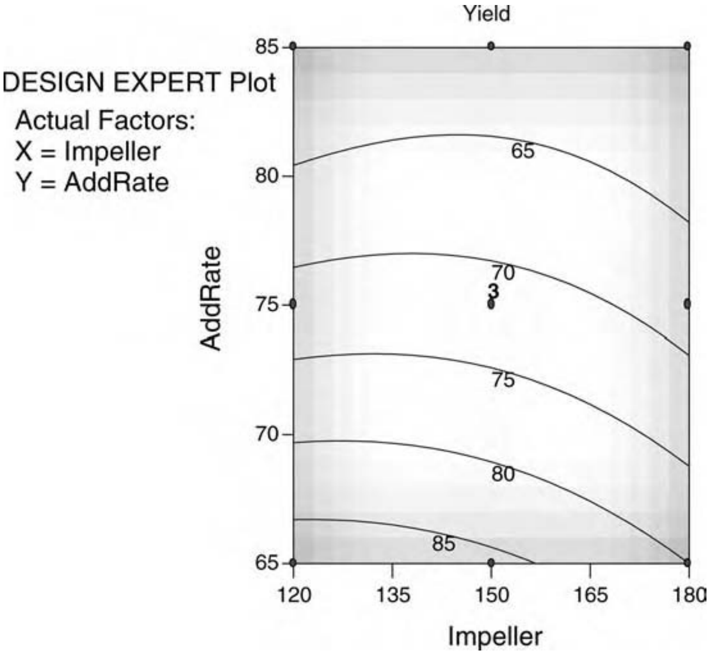


**Figure 7** Granulation milling study—contour plot for yield.

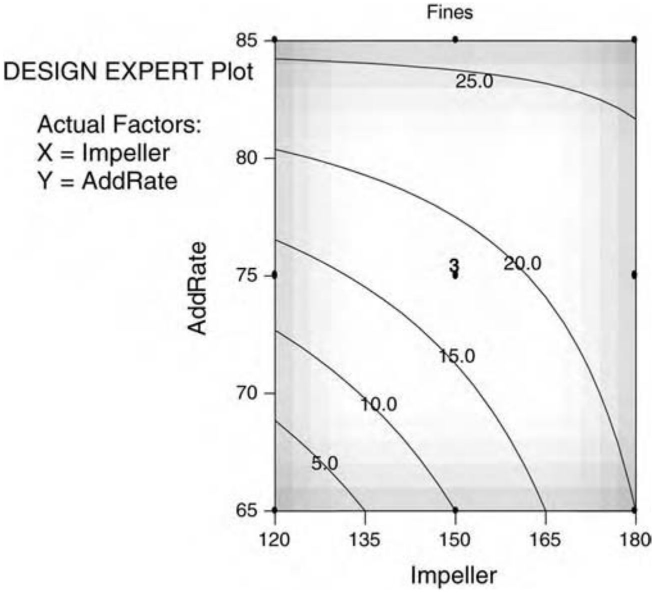


**Figure 8** Granulation milling study—contour plot for fines.

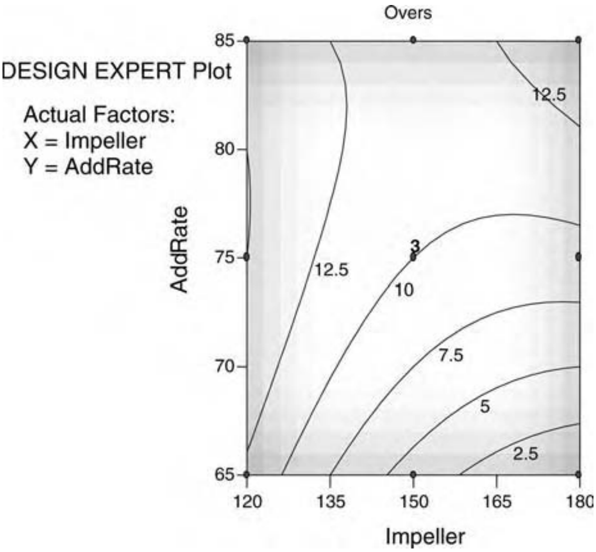


**Figure 9** Granulation milling study—contour plot for overs.

**Table 11** Design and Results for the Coating Study

| Block | Spray rate (g/min) | Product temp (°C) | Atomiz. press. (Bar) | Spray rate coded | Product temp. coded | Atomiz. press. coded | Dissoln. (%) | Resid. solv. (ppm) |
|---|---|---|---|---|---|---|---|---|
| 1 | 650 | 54 | 2.3 | –1 | –1 | –1 | 71.20 | 282.6 |
| 1 | 1050 | 54 | 3.5 | 1 | –1 | 1 | 57.23 | 884.1 |
| 1 | 650 | 60 | 3.5 | –1 | 1 | 1 | 87.32 | 353.5 |
| 1 | 1050 | 60 | 2.3 | 1 | 1 | –1 | 54.78 | 475.4 |
| 1 | 850 | 57 | 2.9 | 0 | 0 | 0 | 71.10 | 490.0 |
| 1 | 850 | 57 | 2.9 | 0 | 0 | 0 | 68.91 | 488.4 |
| 2 | 650 | 54 | 3.5 | –1 | –1 | 1 | 60.70 | 225.9 |
| 2 | 1050 | 54 | 2.3 | 1 | –1 | –1 | 58.22 | 1030.7 |
| 2 | 650 | 60 | 2.3 | –1 | 1 | –1 | 89.74 | 356.7 |
| 2 | 1050 | 60 | 3.5 | 1 | 1 | 1 | 61.27 | 445.8 |
| 2 | 850 | 57 | 2.9 | 0 | 0 | 0 | 69.31 | 494.5 |
| 2 | 850 | 57 | 2.9 | 0 | 0 | 0 | 68.31 | 486.0 |
| 3 | 850 | 57 | 1.9 | 0 | 0 | –1.682 | 68.29 | 529.6 |
| 3 | 850 | 57 | 3.9 | 0 | 0 | 1.682 | 65.09 | 438.2 |
| 3 | 514 | 57 | 2.9 | –1.682 | 0 | 0 | 79.31 | 205.6 |
| 3 | 1186 | 57 | 2.9 | 1.682 | 0 | 0 | 46.55 | 848.2 |
| 3 | 850 | 52 | 2.9 | 0 | –1.682 | 0 | 60.55 | 681.9 |
| 3 | 850 | 62 | 2.9 | 0 | 1.682 | 0 | 79.39 | 358.9 |
| 3 | 850 | 57 | 2.9 | 0 | 0 | 0 | 72.30 | 505.0 |
| 3 | 850 | 57 | 2.9 | 0 | 0 | 0 | 73.24 | 487.0 |

## Case Study 3—Dissolution and Residual Solvent Control in Tablet Coating: A Three-Factor Response Surface Methodology

A tablet-coating investigation was conducted to evaluate the effect of spray rate, inlet air temperature, and atomization air pressure on the six-hour dissolution and residual solvent levels of coated tablets. The goal was to find conditions which maximized the dissolution and minimized the residual solvents of the coated tablets.

This study illustrated the use of a three-factor CCD composed of a $2^3$ factorial design (eight runs), six center points, and six star points, conducted sequentially in three blocks. The sequential design strategy for the 20-run design was summarized in Table 5. As previously discussed, the advantage of this sequential strategy is that an evaluation of the data can be conducted after each block, which can lead to alteration of factor levels, if necessary, without committing to the full design at the outset. The full design is given in Table 11, listing the factor levels (actual and coded units) and the resulting observed responses.

The regression analysis output from MINITAB is listed in Tables 12A and 12B for the two responses, dissolution and solvent, respectively. The regression coefficients are listed for coded and uncoded levels of the factors. The magnitudes of the coded coefficients are more comparable than the uncoded coefficients as the scales of the three factors were standardized to a common scale to make the scales of the coded coefficients equal in magnitude.

The contour plots are shown in Figures 10 and 11. The response surface was a three-dimensional hypersurface in a four-dimensional space, which was hardly able to be plotted. Instead, the response surface for spray rate and air temperature—the two dominating factors—was “sliced” at three levels of the third factor, atomizing air pressure. These were the low, mid, and high factorial levels of 2.3, 2.9, and 3.5 atmospheres (bar) for air pressure.

For dissolution, the maximum values occurred at low spray rates and high temperatures across all values of air pressure. Comparison of the three contour plots across levels of air pressure showed little effect of air pressure on

**Table 12A** Regression Analysis Output for Coating Study Dissolution Response: Response Surface Regression—Dissol vs. AirPress, SprayRate, AirTemp

Estimated regression coefficients for dissol: using data in coded units

| Term | Coef. | SE Coef. | T | P |
|---|---|---|---|---|
| Constant | 70.486 | 1.584 | 44.487 | 0.000 |
| Block1 | 0.231 | 1.267 | 0.182 | 0.860 |
| Block 2 | –0.268 | 1.267 | –0.211 | 0.838 |
| AirPress | –0.937 | 1.050 | –0.892 | 0.398 |
| SprayRat | –7.070 | 1.050 | –6.731 | 0.000 |
| AirTemp | 3.913 | 1.050 | 3.725 | 0.006 |
| AirPress*AirPress | –1.090 | 1.023 | –1.066 | 0.318 |
| SprayRat*SprayRat | -2.419 | 1.023 | –2.365 | 0.046 |
| AirTemp*AirTemp | 0.069 | 1.023 | 0.068 | 0.948 |
| AirPress*SprayRat | 2.303 | 1.373 | 1.678 | 0.132 |
| AirPress*AirTemp | 1.945 | 1.373 | 1.417 | 0.194 |
| SprayRat*AirTemp | –4.070 | 1.373 | –2.965 | 0.018 |
| S = 3.882 | R–Sq = 90.9% | | R–Sq(adj)=78.4% | |

Analysis of variance for dissol

| Source | df | Seq SS | Adj SS | Adj MS | F | P |
|---|---|---|---|---|---|---|
| Blocks | 2 | 0.78 | 0.760 | 0.380 | 0.03 | 0.975 |
| Regression | 9 | 1206.01 | 1206.009 | 134.001 | 8.89 | 0.003 |
| Linear | 3 | 903.85 | 903.847 | 301.282 | 19.99 | 0.000 |
| Square | 3 | 96.97 | 96.966 | 32.322 | 2.14 | 0.173 |
| Interaction | 3 | 205.20 | 205.195 | 68.398 | 4.54 | 0.039 |
| Residual error | 8 | 120.56 | 120.563 | 15.070 | | |
| Lack–of–fit | 5 | 117.22 | 117.223 | 23.445 | 21.06 | 0.015 |
| Pure error | 3 | 3.34 | 3.340 | 1.113 | | |
| Total | 19 | 1327.35 | | | | |

Estimated regression coefficients for dissol: using data in uncoded units

| Term | Coef. |
|---|---|
| Constant | –116.120 |
| Block 1 | 0.230658 |
| Block 2 | –0.267675 |
| AirPress | –61.9031 |
| SprayRat | 0.398463 |
| AirTemp | 3.05722 |
| AirPress*AirPress | –3.02761 |
| SprayRat*SprayRat | –6.04743E–05 |
| AirTemp*AirTemp | 0.00771434 |
| AirPress*SprayRat | 0.0191875 |

(*Continued*)

**Table 12A** Regression Analysis Output for Coating Study Dissolution Response: Response Surface Regression—Dissol vs. AirPress, SprayRate, AirTemp (*Continued*)

Estimated regression coefficients for dissol: using data in coded units

| Term | Coef. |
|---|---|
| AirPress*AirTemp | 1.08056 |
| SprayRat*AirTemp | –0.00678333 |

*Abbreviations*: Adj MS, adjusted mean square; Adj SS, adjusted sum of squares; Coef., estimated coefficient; df, degrees of freedom; F, f statistic; P, p-value; R-Sq, multiple coefficient of determination; R-Sq(adj), adjusted multiple coefficient of determination; S, residual standard deviation; SE Coef., standard error of estimated coefficient; Seq SS, sequential sum of squares; T, t statistic.

**Table 12B** Regression Analysis Output for Coating Study Solvent Response: Response Surface Regression—Solvent vs. AirPress, SprayRate, AirTemp

Estimated regression coefficients for solvent: using data in coded units

| Term | Coef. | SE Coef. | T | P |
|---|---|---|---|---|
| Constant | 491.8 | 4.101 | 119.925 | 0.000 |
| Block 1 | –7.1 | 3.279 | –2.175 | 0.061 |
| Block 2 | 3.8 | 3.279 | 1.159 | 0.280 |
| AirPress | –28.5 | 2.719 | –10.498 | 0.000 |
| SprayRat | 197.5 | 2.719 | 72.661 | 0.000 |
| AirTemp | –97.8 | 2.719 | –35.956 | 0.000 |
| AirPress*AirPress | –3.9 | 2.647 | –1.461 | 0.182 |
| SprayRat*SprayRat | 11.3 | 2.647 | 4.280 | 0.003 |
| AirTemp*AirTemp | 9.0 | 2.647 | 3.412 | 0.009 |
| AirPress*SprayRat | –14.5 | 3.552 | –4.092 | 0.003 |
| AirPress*AirTemp | 21.3 | 3.552 | 5.999 | 0.000 |
| SprayRat*AirTemp | –149.4 | 3.552 | –42.059 | 0.000 |
| S = 10.05 | R-Sq = 99.9% | | R-Sq(adj)=99.8% | |

Analysis of variance for solvent

| Source | df | Seq SS | Adj SS | Adj MS | F | P |
|---|---|---|---|---|---|---|
| Blocks | 2 | 513 | 478 | 239 | 2.37 | 0.156 |
| Regression | 9 | 861,819 | 861,819 | 95,758 | 948.49 | 0.000 |
| Linear | 3 | 674,664 | 674,664 | 224,888 | 2E+03 | 0.000 |
| Square | 3 | 3238 | 3238 | 1079 | 10.69 | 0.004 |
| Interaction | 3 | 183,917 | 183,917 | 61,306 | 607.24 | 0.000 |
| Residual error | 8 | 808 | 808 | 101 | | |
| Lack-of-fit | 5 | 608 | 608 | 122 | 1.83 | 0.328 |
| Pure error | 3 | 199 | 199 | 66 | | |
| Total | 19 | 863,139 | | | | |

(*Continued*)

**Table 12B** Regression Analysis Output for Coating Study Solvent Response: Response Surface Regression—Solvent vs. AirPress, SprayRate, AirTemp (*Continued*)

Estimated regression coefficients for solvent: using data in uncoded units

| Term | Coef. |
|---|---|
| Constant | –5383.33 |
| Block 1 | –7.13220 |
| Block 2 | 3.80113 |
| AirPress | –557.182 |
| SprayRat | 15.0517 |
| AirTemp | 30.3145 |
| AirPress*AirPress | –10.7431 |
| SprayRat*SprayRat | 0.000283289 |
| AirTemp*AirTemp | 1.00378 |
| AirPress*SprayRat | –0.121146 |
| AirPress*AirTemp | 11.8403 |
| SprayRat*AirTemp | –0.249021 |

*Abbreviations*: Adj MS, adjusted mean square; Adj SS, adjusted sum of squares; Coef., estimated coefficient; df, degrees of freedom; F, f statistic; P, p-value; R-Sq, multiple coefficient of determination; R-Sq(adj), adjusted multiple coefficient of determination; S, residual standard deviation; SE Coef., standard error of estimated coefficient; Seq SS, sequential sum of squares; T, t statistic.

dissolution. For residual solvents, the minimum values occurred at low spray rates and low air temperatures, with the air temperature interacting with the air pressure. High air pressure gave lower solvent levels at the low air temperatures. Thus, there would be a tradeoff of higher dissolution against lower residual solvent.

When there are a larger number of responses, the graphical approach to analysis with regard to meeting specifications for all responses can be daunting, if not impossible. An alternative approach is to use a systematic grid search over the factor space. At every point on the grid, each response is examined for conformance to the specifications (yes or no). A grid point that meets specifications for all responses is termed a “hit.” Finding the minimum cost “hit” is the objective. If there are no “hits,” then the specification windows must be relaxed, if possible.

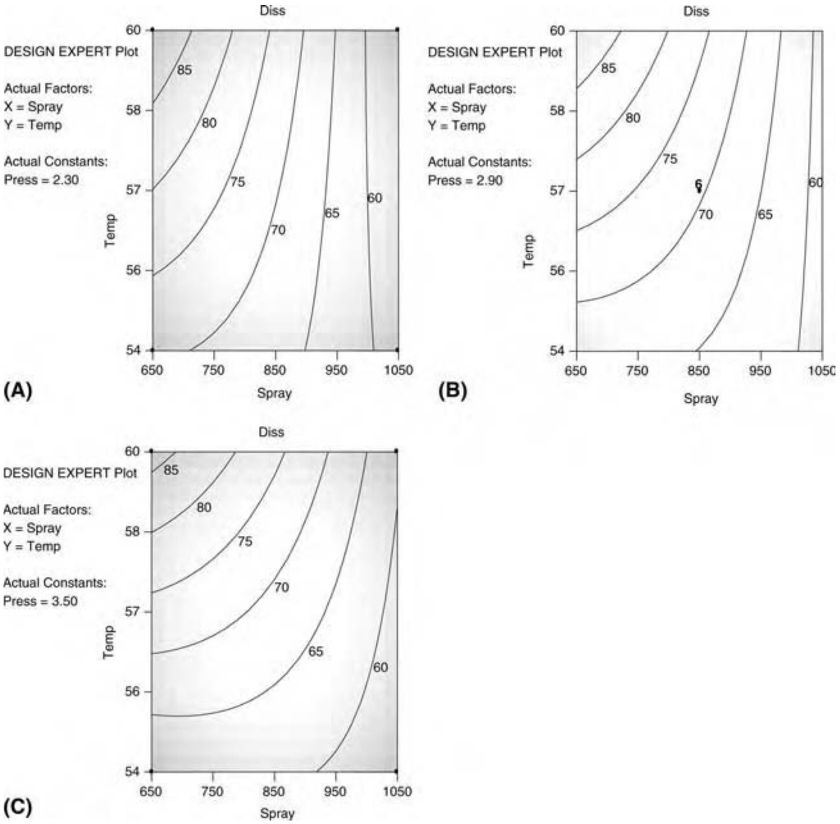


**Figure 10** (**A**) Tablet coating study—contour plot for dissolution spray rate and air temperature. Atomization air pressure, 2.3 Bar. (**B**) Tablet coating study—contour plot for dissolution spray rate and air temperature. Atomization air pressure, 2.9 Bar. (**C**) Tablet coating study—contour plot for dissolution spray rate and air temperature. Atomization air pressure 3.5 Bar.

In summary, RSM is a useful technique for finding the optimum conditions for one or more responses over up to about five factors. The types of experimental designs often used for RSM are the CCD and Box–Behnken designs. The response surface can be well-described by a second-order polynomial model, and thus can be used to readily find the optimum conditions for a single response or to perform a tradeoff analysis among two or more responses.

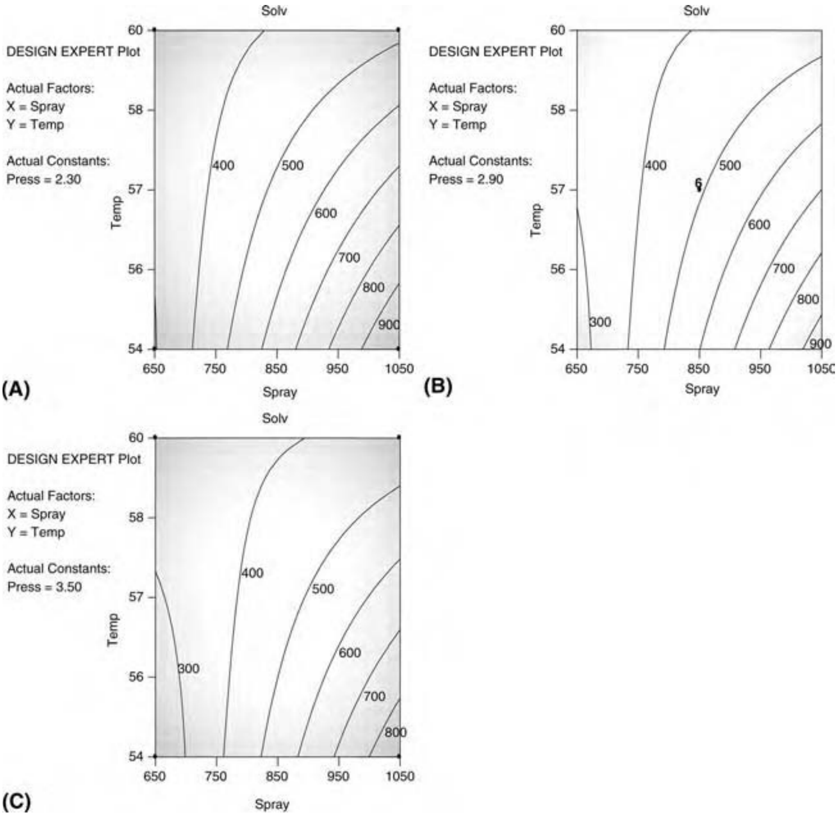


**Figure 11** (**A**) Tablet coating study—contour plot for residual solvent spray rate and air temperature. Atomization air pressure 2.3 Bar. (**B**) Tablet coating study—contour plot for residual solvent spray rate and air temperature. Atomization air pressure 2.9 Bar. (**C**) Tablet coating study—contour plot for residual solvent spray rate and air temperature. Atomization air pressure 3.5 Bar.

## REFERENCES

1. Davies OL. The Design and Analysis of Industrial Experiments. New York: Hafner 1960.

2. Box GEP, Hunter WG, Hunter JS. Statistics for Experimenters. New York: Wiley 1978.

3. Myers RH, Montgomery DC. Response Surface Methodology. New York: Wiley 1995.

7

# The Role of Designed Experiments in Developing and Validating Control Plans

**WAYNE A. TAYLOR**

Taylor Enterprises, Inc.
Libertyville, Illinois, U.S.A.

## INTRODUCTION

Process validation is "establishing documented evidence which provides a high degree of assurance that a specific process will consistently produce a product meeting its predetermined specifications and quality attributes" (1). Being able to validate a process first requires that a process be developed that meets the customer requirements. This development process requires taking the external customer requirements and translating them into internal requirements for the manufacturing parameters, materials, procedures, and environment that ensure that the customer requirements are met. It also requires establishing controls to ensure that these internal requirements continue to be met.

Designed experiments are a key tool for performing this specification translation process and helping to establish such controls. However, designed experiments are not the only tool required to accomplish this task. We will also explore other tools, such as tolerance analysis, robust design, capability studies, and Failure Modes and Effects Analysis (FMEA), to see how to combine these tools into an effective system for validation.

The ideal is to set operating windows on the internal parameters, which ensure that the external requirements are met. Achieving this ideal requires identifying all the internal parameters affecting the external requirements and that these internal parameters be adequately controlled. This is often not the case.

For example, having validated an injection molding process and specifying settings for all the machine parameters, the operator may find themselves in a predicament. A batch of resin has been received with a higher melt index. In the past, the operator would have increased the barrel temperature to compensate. However, as a part of validation, the temperature range has been restricted so that they can no longer make this adjustment. They go to their supervisor and ask: "What do you want me to do? Make good product by running outside the validated operating window or run inside the validated operating window and make bad product?" The problem is that not all the key parameters were included when setting the operating windows. Besides machine parameters, operating windows need to be set on material properties, environmental factors and operator effects. However, setting an operating window for the melt index of the material does not necessarily solve the problem if the suppliers cannot meet this operating window. This might require purchasing a more expensive grade of material, pricing the product out of the market. This chapter addresses incorporating adjustment mechanisms within the framework of a process validation.

As a second example, latex, used to produce latex gloves, is a natural ingredient produced from tree sap, which varies in consistency with weather conditions and geographic location. Each batch of latex must be analyzed to determine its properties and then adjusted to get a more consistent performance. This is an example of a feed-forward control mechanism. One cannot set an operating window on the amount of solids added because it depends on the batch of latex. One cannot set an operating window on the percent solids in the latex initially because nature cannot be controlled. However, a control mechanism can still be established that ensures that the product will consistently meet requirements. Again adjustment mechanisms are required.

This chapter focuses on validating control plans. The Global Harmonization Task Force (GHTF) guideline (2) states:

- "One output of process validation is the development of a control plan";

- "The final phase of validation requires demonstrating this control plan works."

The control plans considered include feed-forward and feedback mechanisms. Designed experiments can provide the detailed understanding required to establish more complex control mechanisms where they are needed.

## THE CONTROL PLAN

A control plan is the sum of the procedures and equipment used to ensure that the internal requirements are met. It includes control charts, sampling plans, 100% inspection, feed-forward/feedback mechanisms, and mistake-proofing techniques/devices. Some items may be performed by equipment like automatic controllers, and some are performed by operators and checkers. There may not be a single document called a control plan. Instead, these controls might be spread across a number of documents including a statistical process control plan, an inspection plan, an operator manual, and various other standard operating procedure and specifications.

This control plan is an integral part of the process. One cannot validate a process without first specifying how the process is to be operated. As a part of validation, we want to prove that the control plan works.

The control plan is designed to prevent defects, that is, product that does not meet its "predetermined specifications and quality attributes." Some defects result from errors. For example, an operator might forget to insert a part, might incorrectly load a part into a welder or perform an operation they were not supposed to. The tool for addressing this type of defect is mistake proofing. There are several mistake-proofing strategies, including:

- Elimination—make it impossible for the defect to occur;
- Facilitation—make it easier to do it right;
- Replacement—replace less reliable processes with more reliable processes;

- Flagging—make mistakes more visible so they are detected and removed;
- Redundant—add redundancy so a single mistake does not cause a product to fail;
- Fail-safe—lessen impact of mistake should it occur.

Elimination is generally the preferred strategy, but not always possible. Designed experiments can sometimes be used as a part of the facilitation strategy to identify conditions that affect the rate of mistakes and to identify the conditions that minimize the incidence of mistakes.

Before mistake proofing can be applied, potential mistakes must first be identified. FMEA can be used for this purpose. It identifies different failure modes along with their potential causes and consequences. For each potential failure mode, a risk evaluation is performed based on the likelihood of a defect to occur, the likelihood of it being detected, and the severity of its consequences. One of the columns in an FMEA is titled "Control Plan." This column must be filled out before performing the risk assessment. Both the likelihood of occurrence and likelihood of detection are affected by the controls that are currently in place.

FMEA evaluate the current control plan and identify high-risk areas that require additional controls. As high-risk items are identified and dealt with, the control plan evolves. Once the risks are all at acceptable levels, the control plan is believed adequate and it is time to proceed with the worst-case and performance qualification (PQ) testing phases of validation.

Some of the high-risk items can be dealt with by mistake proofing. However, other failure modes will be found that are not mistakes but, instead, centering and variation issues. For example, a seal could leak due to inadequate sealing, which is affected by dozens of factors, including seal time, seal temperature, pressure, material thickness, material temperature, room temperature, and so on. Resolving this issue requires identifying the different factors affecting seal strength, establishing targets and operating windows for these factors, and establishing controls to maintain these factors inside their respective operating windows. It might also require establishing more complex controls like feedback and feed-forward. This type of situation is the topic of the next section.

## THE SPECIFICATION TRANSLATION PROCESS

For measurable characteristics, the process of translating customer requirements into internal requirements can be pictured in terms of an input/output model. Take as an example the customer requirement that the dosage of a vial of drug contain ±15% of the labeled dosage of 10 mg. Dosage is the external requirement and will be referred to as an output variable. A vial's dosage depends on the fill volume of the vial and the concentration of the solution. These are the internal requirements and will be referred to as input variables or factors. The input/output (I/O) system is shown in Figure 1.

Translating specifications for external requirements into internal requirements requires the following five items of information:

1. Identify the critical output variables;
2. Develop measurements for these output variables and establish specifications based on customer need;
3. Identify the critical input variables affecting the outputs;
4. Model the effect of the critical inputs on the critical outputs;
5. Determine manufacturing and supplier capabilities to control the critical input variables.

The goal is to set specifications on the critical inputs that ensure that the output specifications are met and can be met

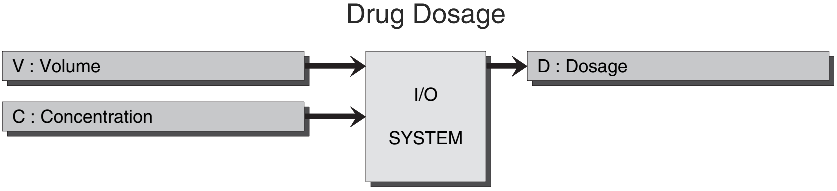


**Figure 1** Input/output system for drug dosage.

in manufacturing. While gathering all the aforementioned information takes time and resources, one must ask how a design can be completed if any of these items are missing. How can one design if one does not know the critical outputs; one does not know what the requirements are; one does not know the critical inputs/factors that affect the output; one does not know how the inputs affect the output; one does not understand the ability to control the critical inputs. With this knowledge, the design can be optimized. Without any one of these items, the design process degenerates into one of trial and error resulting in marginal performance and manufacturability issues.

Different tools are required for each item. Items 1 and 2 require the customer's input (voice of the customer). The tools include data-gathering tools, like focus groups and surveys, and statistical analysis tools, like conjoint and regression analysis. Measurements may have to be developed requiring the use of gage studies.

Having identified the critical outputs and their specifications, we must determine the critical inputs (item 3). A type of design of experiment (DOE), called a screening experiment (fractional factorial), can be used to efficiently sort through a large number of potential inputs to identify those that are critical. A case study is presented illustrating the use of screening experiments.

Subsequently, we need to understand how the critical inputs affect the critical outputs (item 4). A second type of designed experiment, called response surface methodology (RSM), is used to accomplish this task. Sometimes we are fortunate and know the equation in advance. For example, the equation for dosage above is $D = V \times C$. However, when the equation is not known, a DOE can be used to empirically fit a model. A response surface study is also presented as part of this case study.

There is more than one way to set the specifications for the critical inputs. Tightening the specification of one input variable may allow us to open up the specification on another. How do we decide which set of specifications are best? The answer is based on the ability to meet the specifications.

This requires the gathering of capability data and life-testing data from the plants and suppliers to determine which specifications can be easily met (item 5, voice of manufacturing).

Having gathered all this information, we are not done. To come up with the final set of specifications, tolerance analysis and design optimization methods are required. In optimizing the design, care should be taken to make the design robust to the variation of the inputs to help open up the tolerances for the inputs. Hopefully, the result will be a set of specifications for the inputs that ensure the output specifications are met and are manufacturable. Sometimes, specifications for the inputs must be tighter than desired, requiring alternate suppliers and processes or 100% inspection. Once the specifications for the critical inputs are set, the last step is to specify the controls required to ensure that these specifications continue to be met.

## HEAT SEALER CASE STUDY

To illustrate the use of designed experiments in this process, a case study is presented involving a packaging sealer (3). Designed experiments in conjunction with other tools are used to create a control plan that is then validated. The DOE results are also used to identify worst-case conditions for worst-case testing, and to help select sample sizes for worst-case and final PQ testing.

The case study will be limited to forming the top seal of a plastic pouch used to protect the product. This pouch serves as both a moisture and sterility barrier, so it is considered critical to the safety of the product inside. The seal is torn open by the customer to remove the product. There is both a lower limit on strength to ensure seal integrity and an upper limit on strength to ensure that the customer can comfortably open the bag.

Two critical outputs have been identified. The first is seal strength and the second is visible discoloration of the seal called "Seal Burn." A measurement method has been developed for measuring seal strength. It involves cutting a 1-inch wide strip of the seal and measuring the force required to pull the seal apart in a tensile tester. Further testing with the customer has identified the upper specification limit should be set to 32 lb. Accelerated life testing using elevated temperatures and shaking the pouch identified a lower limit of 20 lb. This results in a specification for seal strength of 26 ± 6 lb. Limit samples have also been established based on customer input for the amount of discoloration allowed before a bag is considered to have seal burn. A pass/fail characteristic like this is referred to as attribute data. This completes items 1 and 2.

The next step is to identify the critical inputs affecting seal strength and seal burn (item 3). This might start with a brainstorming session with the operators and engineers to identify the possible inputs. It should also involve checking with the equipment and material vendors and reviewing similar validations. Based on the aforementioned, the following candidate input variables were identified:

| | |
|---|---|
| HB | Hot bar temperature |
| CB | Cold bar temperature |
| DT | Dwell time |
| P | Pressure |
| CA | Cooling air pressure |
| TH | Material thickness |
| MT | Material temperature |
| RT | Room temperature |

The heat sealer works by pinching the material to be sealed between two bars. The top bar, called the hot bar, provides heat to melt the plastic material, causing it to flow together to form the seal. The top bar also moves up and down to allow the material to be moved. The bottom bar is stationary and has cooling water running through it, allowing its temperature to be controlled.

When the top bar comes down to make contact with the material, it is lowered until it exerts a preset pressure on the material. It is then held there for a preset time. Before moving the material, cooling air is blown on the seal to facilitate hardening.

When determining input variables/factors to include, one must not restrict oneself to just machine parameters. Material properties, environmental factors, operator effects, and procedures should also be considered. Otherwise, you might suffer the fate of the injection molding operator discussed earlier, unable to make compensating adjustments as unidentified critical inputs vary.

## SCREENING EXPERIMENT

Just because an input might affect the output does not mean that it does. The next step is to reduce the larger list of candidate input variables into a smaller list of the critical inputs. A type of designed experiment called a screening experiment (3) can be used for this purpose. They are also commonly referred to as fractional factorial designs.

While the primary purpose of the screening experiment is to identify the critical inputs, a secondary goal is to determine what type of model to fit to the inputs when a follow-up response surface study is run. This will affect our choice of what design to run.

Having identified the inputs to include in the study, a range must be selected for each input. The ranges selected should be as wide as possible in order to magnify the effects of the different variables. In Figure 2, it is shown how selecting

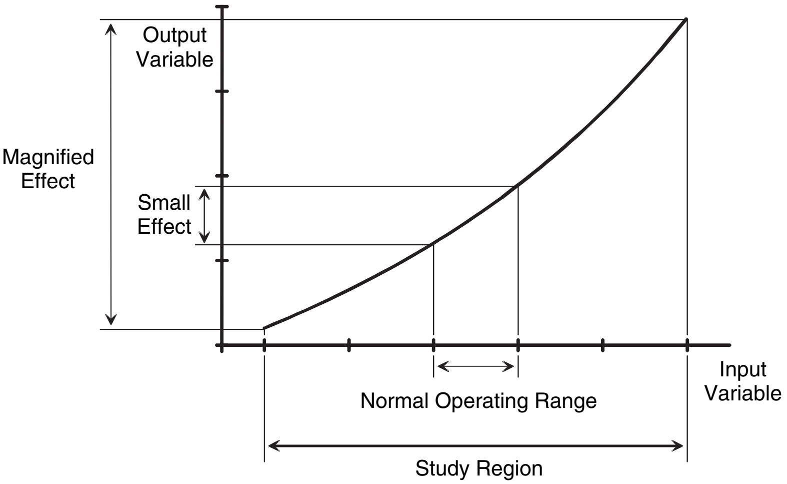


**Figure 2** Wide study regions magnify the effects of the inputs on the output.

a study region five times the normal operating window used in production magnifies the effect of the input, making it easier to detect the effect. Production operating windows are generally selected to make the effect of the input small. To see these effects more clearly, broad study regions are needed. Studying the process over a region five times larger than the normal operating window is similar to putting your process under a microscope and flipping the lens to 5X power. Suddenly effects become visible that could not previously be seen.

The tendency is to set the study regions too narrow. However, it is also possible to go too far the other way. One does not want to set the study region so extreme that units are so poorly formed to prevent proper measurement or so extreme so as to cause damage to the equipment. Sometimes, small longitudinal studies are run first, adjusting a couple of the more important variables up and down to determine the limits of operability.

For the eight inputs, the following study regions were selected:

| | | |
|---|---|---|
| HB | Hot bar temperature | 150–200°F |
| CB | Cold bar temperature | 80–120°F |
| DT | Dwell time | 0.5–1.0 seconds |
| P | Pressure | 50–150 lbs |
| CA | Cooling air pressure | 0–30 lbs |
| TH | Material thickness | 14–15 mils |
| MT | Material temperature | 70–110°F |
| RT | Room temperature | 70–80°F |

This provides the details needed to complete the design of the screening experiment. It was decided to run a screening experiment with the 22 trials shown in Table 1. For each of the 22 trials, the process is set to the specified conditions, and time is allowed for the machine to reach the set conditions and then a series of units are produced. For each trial, five units were randomly selected and tested for seal strength and 50 units were inspected for burns. The resulting data is summarized in Table 1. For each trial, the average and standard deviation of the five units is shown along with the number of units with burns.

**Table 1** Trials and Data for Screening Experiment

| Trial | HB | CB | DT | P | RT | CA | TH | MT | SS-Ave | SS-SD | No. of burns | Order |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 150 | 80 | 0.5 | 50 | 70 | 0 | 14 | 70 | 9.80 | 2.376 | 0 | 3 |
| 2 | 150 | 80 | 0.5 | 150 | 80 | 30 | 14 | 110 | 16.99 | 4.428 | 0 | 7 |
| 3 | 150 | 80 | 1 | 50 | 80 | 30 | 15 | 70 | 17.22 | 1.146 | 0 | 8 |
| 4 | 150 | 80 | 1 | 150 | 70 | 0 | 15 | 110 | 22.09 | 2.779 | 0 | 17 |
| 5 | 150 | 120 | 0.5 | 50 | 80 | 0 | 15 | 110 | 13.67 | 3.372 | 0 | 10 |
| 6 | 150 | 120 | 0.5 | 150 | 70 | 30 | 15 | 70 | 12.55 | 2.898 | 0 | 15 |
| 7 | 150 | 120 | 1 | 50 | 70 | 30 | 14 | 110 | 19.73 | 1.572 | 2 | 18 |
| 8 | 150 | 120 | 1 | 150 | 80 | 0 | 14 | 70 | 21.84 | 3.135 | 0 | 4 |
| 9 | 200 | 80 | 0.5 | 50 | 70 | 30 | 15 | 110 | 26.55 | 1.325 | 0 | 2 |
| 10 | 200 | 80 | 0.5 | 150 | 80 | 0 | 15 | 70 | 29.15 | 1.499 | 0 | 14 |
| 11 | 200 | 80 | 1 | 50 | 80 | 0 | 14 | 110 | 8.64 | 3.234 | 2 | 9 |
| 12 | 200 | 80 | 1 | 150 | 70 | 30 | 14 | 70 | 11.95 | 3.892 | 4 | 5 |
| 13 | 200 | 120 | 0.5 | 50 | 80 | 30 | 14 | 70 | 23.67 | 1.467 | 0 | 1 |
| 14 | 200 | 120 | 0.5 | 150 | 70 | 0 | 14 | 110 | 30.91 | 1.580 | 0 | 12 |
| 15 | 200 | 120 | 1 | 50 | 70 | 0 | 15 | 70 | 9.90 | 8.100 | 4 | 16 |
| 16 | 200 | 120 | 1 | 150 | 80 | 30 | 15 | 110 | 17.61 | 5.495 | 5 | 6 |
| C1 | 175 | 100 | 0.75 | 100 | 75 | 15 | 14.5 | 90 | 28.46 | 1.980 | 0 | 11 |
| C2 | 175 | 100 | 0.75 | 100 | 75 | 15 | 14.5 | 90 | 29.28 | 0.923 | 0 | 13 |
| R1 | 150 | 80 | 0.5 | 50 | 70 | 0 | 14 | 70 | 12.72 | 4.171 | 0 | 21 |
| R8 | 150 | 120 | 1 | 150 | 80 | 0 | 14 | 70 | 22.06 | 1.610 | 0 | 22 |
| R9 | 200 | 80 | 0.5 | 50 | 70 | 30 | 15 | 110 | 25.99 | 1.860 | 0 | 20 |
| R13 | 200 | 120 | 0.5 | 50 | 80 | 30 | 14 | 70 | 26.34 | 0.762 | 0 | 19 |

*Abbreviations*: CA, cooling air pressure; CB, cold bar temperature; DT, dwell time; HB, hot bar temperature; MT, material temperature; P, pressure; RT, room temperature; SD, standard deviation; SS, seal strength; TH, material thickness.

The trials shown in Table 1 started with a resolution IV fractional factorial design allowing independent estimates of all eight factors/inputs, unconfounded both with each other and with any two-way interactions. The base design consists of the 16 unique trials labeled 1–16. To the base design, two center points were added (C1 and C2). This allows the presence of quadratic effects to be detected. Finally, four trials selected at random were run a second time (R1, R8, R9, R13). This allows an estimate of the variation to be made for testing the effects so that one does not have to use a normal probability plot and rely on only a few effects being significant. The total number of trials is 22 = 16 base trials + 2 center points + 4 repeated trials. Table 1 shows the trials in standard order with the added trials at the end. The trials were actually run in the random order shown in the last column.

This design was selected, not only to determine the critical inputs, but also to help determine the model to fit to the data in a follow-up response surface study. Center points were added to determine if quadratic effects are needed in the model. Also of concern are interactions. There are 28 possible two-way interactions that might exist between these eight inputs. These two-way interactions are confounded together and are not estimable separately. However, the aforementioned design groups these interactions into seven distinct groups of four interactions each. Each of these groups can be tested for significance. As many of these groups are expected to be nonsignificant, it may be possible to eliminate the majority of these interactions.

The result of the analysis of the seal strength averages is shown in Table 2. A term is considered significant if its *p*-value is 0.05 or below. Those terms that are significant are shown in italic. The *p*-value represents the probability the observed effect for the term could result from the noise in the data alone. For each significant effect, it can be stated with 95% confidence that the term's effect is not zero. All the terms of a quadratic polynomial are shown in the effects table, including:

**Table 2** Effects Table for Seal Strength Average—Screening Experiment

| Term | Effect | *P*-value |
|---|---|---|
| ***Intercept*** | | ***0.000*** |
| ***Hot bar temperature (HB)*** | ***2.997*** | ***0.004*** |
| Cold bar temperature (CB) | 0.969 | 0.170 |
| ***Dwell time (DT)*** | ***–4.589*** | ***0.001*** |
| ***Pressure (P)*** | ***3.938*** | ***0.001*** |
| Room temperature (RT) | 0.697 | 0.301 |
| Cooling air pressure (CA) | –0.031 | 0.962 |
| Material thickness (TH) | 0.253 | 0.693 |
| ***Material temperature (MT)*** | ***2.116*** | ***0.017*** |
| HB*CB, DT*TH, P*MT, RT*CA | 0.884 | 0.203 |
| ***HB*DT, CB*TH, P*CA, RT*MT*** | ***–11.219*** | ***0.000*** |
| HB*P, CB*MT, DT*CA, RT*TH | 1.013 | 0.154 |
| HB*RT, CB*CA, DT*MT, P*TH | –0.353 | 0.584 |
| HB*CA, CB*RT, DT*P, TH*MT | 0.589 | 0.374 |
| HB*TH, CB*DT, P*RT, CA*MT | 1.353 | 0.075 |
| HB*MT, CB*P, DT*RT, CA*TH | –0.259 | 0.686 |
| ***TH^2, CB^2, DT^2, P^2, RT^2, CA^2, TH^2, MT^2*** | | ***0.000*** |

1 Intercept term
8 Linear terms (represented by the name of the variable)
8 Quadratic terms (TH^2, CB^2, etc.)
28 Interaction terms (HB*DT, CB*TH, etc.)

The linear terms are used to determine the critical inputs. Four of the inputs are significant: HB, DT, P, and MT. These are the critical inputs.

The interaction and quadratic terms are used to help determine what model to fit to the data. When multiple terms appear of the same line, it means they are confounded together. Their individual effects cannot be separated. If the row tests significant, it can be concluded that one or more of the terms is significant. If the row tests nonsignificant, it is usually assumed that all terms are nonsignificant. Six of the interaction rows test nonsignificant allowing the 24 interactions on these rows to be eliminated. One interaction row tests significant. These four interactions cannot be eliminated based on the data. However, reason tells us that the HB*DT interaction is the only interaction on this row between two critical variables, and thus most likely to be responsible for the row testing significant. The last row contains all eight quadratic terms. The fact that it tests significant indicates that the quadratic terms for all critical inputs should be included in the model to be fit.

The effects column represents the amount that the seal strength average changes as a result of adjusting the inputs from its low setting to its high setting. The estimated effect of the HD*DT interaction row is –11.219. It is much larger than any other effect. Understanding this interaction effect is critical for the understanding of the heat sealer.

The standard deviation can also be analyzed. However, care must be taken because the standard deviations of each trial are of five units over a very limited period of time and from a single roll of material. They might underestimate the variation experienced in full-scale production. We will return to this issue later.

When analyzing seal burns, the terms cannot be tested for significance because all the repeated trials had zero defects resulting in an error estimate of zero. However, a normal probability plot of the effects can be performed. Nonsignificant effects should form a line, whereas significant effects should fall off of this line. The result is shown in Figure 3. Both temperature hot bar and dwell time test significant. The group of interactions associated with the HB*DT interaction also tests significant.

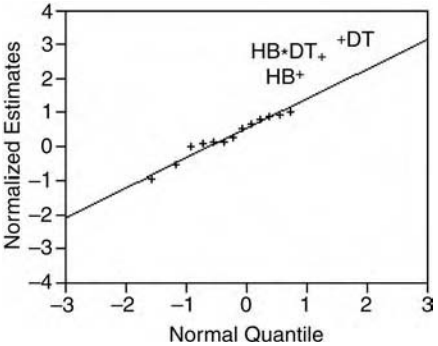


**Figure 3** Normal probability plot of effects on burn.

To summarize, the screening experiment has identified four critical inputs: HB, DT, P, and MT. It also indicates that the HB*DT interaction group and quadratic effects should be included in the model. We are now ready to develop a model for the process.

## RESPONSE SURFACE METHOD STUDY

Now that the screening experiment has determined the critical inputs, a response surface study can be run to help understand the relationship between these critical inputs and the outputs. The response surface study will provide plots of the effects of the critical inputs as well as an equation.

The data already collected during the screening experiment can be reused. However, the response surface study also requires additional data to be collected. In order to determine which quadratic effects exists, the six additional trials given in Table 3 must be run. The resulting data is also shown. These trials were selected using the D-optimal design method (4) with the trials from the screening experiment used as the starting design. The other four inputs that did not test significant have been eliminated from the study. As they do not affect either output, they were set at low-cost settings. For example, cooling air had no benefit so was turned off. This is an often unrecognized benefit of DOE.

The significant effects for the seal strength average are given in italic in Table 4. In addition to the effects previously

**Table 3** Additional Trials and Data for Response Surface Study

| Trial | HB | DT | P | RT | SS-Ave | SS-SD | No. of burns |
|---|---|---|---|---|---|---|---|
| 23 | 200 | 0.75 | 150 | 90 | 28.08 | 3.698 | 0 |
| 24 | 150 | 0.5 | 100 | 90 | 11.89 | 6.682 | 0 |
| 25 | 150 | 0.75 | 100 | 70 | 18.48 | 1.494 | 0 |
| 26 | 175 | 1 | 100 | 110 | 21.38 | 4.524 | 1 |
| 27 | 175 | 0.75 | 150 | 110 | 31.44 | 0.406 | 1 |
| 28 | 175 | 1 | 50 | 90 | 19.82 | 4.779 | 1 |

*Abbreviations*: HB, hot bar temperature; DT, dwell time; P, pressure; RT, room temperature; SS, seal strength; SD, standard deviation.

**Table 4** Analysis Table for Seal Strength Average—R/S Study

| Term | Coeff | $P$-value |
|---|---|---|
| ***Intercept*** | ***28.482*** | ***0.000*** |
| Block | 0.388 | 0.339 |
| ***Hot bar temperature (HB)*** | ***1.491*** | ***0.000*** |
| ***Dwell time (DT)*** | ***–2.241*** | ***0.000*** |
| ***Pressure (P)*** | ***2.030*** | ***0.000*** |
| ***Material temperature (MT)*** | ***0.986*** | ***0.005*** |
| ***HB*DT*** | ***–5.604*** | ***0.000*** |
| ***HB^2*** | ***–5.818*** | ***0.000*** |
| ***DT^2*** | ***–5.213*** | ***0.000*** |
| P^2 | –1.588 | 0.109 |
| MT^2 | –0.940 | 0.319 |

found, both dwell time and hot bar temperature have quadratic effects. The block effect tests whether a change or shift in the process occurred between the two sets of data. None was detected.

Using a regression analysis involving only those effects testing significant in Table 4 results in the following equation:

$$\begin{aligned}\text{Seal strength} = {} & -409.2736 + 3.7514\ \text{HB} + 263.0555\ \text{DT} \\ & + 0.0408\ \text{P} + 0.0482\ \text{MT} - 0.0086\ \text{HB}^2 \\ & - 75.7987\ \text{DT}^2 - 0.9046\ \text{HB DT}\end{aligned}$$

A couple of additional checks are performed for the adequacy of the model. The coefficient of determination, or $R^2$ value, is 96.8%, which is very good. A lack-of-fit test is performed comparing the error in the estimated values at each data point with the estimated noise obtained from the repeated trials. The lack-of-fit test passes, indicating that the model's fit to the data is within the accuracy expected based on the data's noise.

A model has now been developed for the seal strength average. A total of 28 trials were required. The strategy used, resolution IV screening experiment followed by design augmentation, was selected in order to minimize the number of trials required. Some alternative approaches and the corresponding number of trials are as follows:

Response surface study including all eight candidate inputs—82 trials

Response surface study including just four critical inputs—26 trials (requires guessing right four critical inputs)

The approach used is as efficient as running a response surface study on just the four critical inputs without the risk of having to guess the correct four critical inputs.

The seal strength standard deviation can be modeled as well. The significant effects are given in *italics* in Table 5. The logs of the standard deviations were analyzed, as is standard practice since the distribution of the standard deviation is right skewed.

Using a regression analysis involving only those effects testing significant in Table 5 results in the following equation:

$$\begin{aligned}\text{Log (seal strength SD)} = {} & 24.5885 - 0.3360\ \text{HB} - 33.5870\ \text{DT} \\ & + 0.3847\ \text{MT} + 0.0008\ \text{HB}^2 \\ & + 14.4755\ \text{DT}^2 - 0.0021\ \text{MT}^2 \\ & + 0.0730\ \text{HB DT}\end{aligned}$$

HB and MT are included even though they did not test significant as HB*DT and MT$^2$ are included. The $R^2$ value is 73.98%

**Table 5** Analysis Table for Log of Seal Strength Standard Deviation—R/S Study

| Term | Coeff | *P*-value |
|---|---|---|
| Intercept | 0.40886 | 0.100 |
| Block | –0.10737 | 0.363 |
| Hot bar temperature (HB) | 0.05190 | 0.574 |
| ***Dwell time (DT)*** | ***0.21439*** | ***0.030*** |
| Pressure (P) | 0.08235 | 0.376 |
| Material temperature (MT) | 0.04454 | 0.628 |
| ***HB*DT*** | ***0.44002*** | ***0.000*** |
| ***HB^2*** | ***0.59853*** | ***0.050*** |
| ***DT^2*** | ***1.10414*** | ***0.001*** |
| P^2 | 0.31131 | 0.271 |
| ***MT^2*** | ***–0.73999*** | ***0.013*** |

which is OK. The model also passes the lack-of-fit test. Again, care must be taken because the standard deviations might underestimate the variation experienced in production.

The significant effects for the burn are given in italic in Table 6. The arcsine of the fraction defective is analyzed. Only HB and DT have been identified as critical inputs and were included in the model. The above model has an R-squared value of 79.6% and passes the lack of fit test. Removing the two quadratic terms, which both tested non-significant, causes the lack of fit test to fail. Therefore, it was decided to leave the two quadratic terms in the fitted equation:

$$\begin{aligned}\text{Arcsine (Burn FD)} = {} & 0.734053 - 0.006676\ \text{HB} \\ & - 0.619195\ \text{DT} + 0.000015\ \text{HB}^2 \\ & + 0.158355\ \text{DT}^2 + 0.002629\ \text{HB DT}\end{aligned}$$

Figure 4 contains a contour plot of the previous equation. Percent burn is considered a major defect for which it is desired to be below 1% defective. From the contour plot, we can conclude that if DT is maintained below 0.7 seconds, the percent burns will be below 1%. This is true for all values of HB, P, and MT within the region of study.

We can establish an operating window for percent burn as shown in Table 7. In general, this is the way attribute characteristics will be handled. We construct a model, and then determine the widest possible operating window. We will restrict ourselves to this operating window when dealing with other measurable characteristics like seal strength. We could

**Table 6** Analysis Table for Arcsine of Burn Fraction Defective—R/S Study

| Term | Coeff | *P*-value |
|---|---|---|
| Intercept | 0.001492 | 0.845 |
| Block | 0.002955 | 0.458 |
| ***Hot bar temperature (HB)*** | ***0.014980*** | ***0.000*** |
| ***Dwell time (DT)*** | ***0.019725*** | ***0.000*** |
| ***HB*DT*** | ***0.016721*** | ***0.000*** |
| HB^2 | 0.007748 | 0.394 |
| DT^2 | 0.008023 | 0.368 |

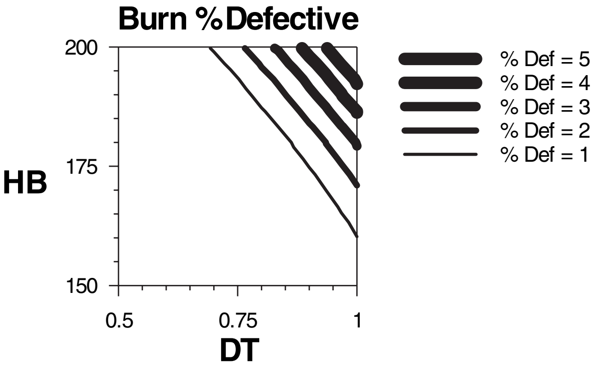


**Figure 4** Contour plot of burn percent defective. *Abbreviations*: DT, dwell time; HB, hot bar temperature.

follow a similar approach with the seal strength average and standard deviation. The specification limits for seal strength are 20 to 32 lb. To achieve a high-quality product, we would like the standard deviation to be no more than one-twelfth the width of this interval, resulting in a maximum standard deviation of 1 lb. We want the seal strength average to be close to 26 lb. Plotting contour plots for both the seal strength average and standard deviation will identify regions were the different requirements are met. Hopefully, a region will be determined that satisfies all the requirements.

This approach has several distinct disadvantages:

- It assumes that the equation for the standard deviation correctly predicts the variation that will be experienced in production. In many cases, the equation will underestimate the full range of variation experienced in manufacturing.

**Table 7** Operating Window for Percent Burn

| Input | Operating window | Worst-case condition |
|---|---|---|
| Hot bar temperature | 150 to 200°F | 200°F |
| Dwell time | 0.5– 0.7 sec | 0.7 sec |
| Pressure | 50 to 150 lbs | Any value |
| Material temperature | 70 to 110°F | Any value |

- It assumes that all critical inputs have been identified and are controlled. Because operating windows are established for all critical inputs, the ability to make compensating adjustments is limited.
- It does not consider manufacturing capability in setting operating windows for the critical inputs. The resulting design may not be manufacturable.
- When there are more than two critical inputs, it is confusing to overlay contour plots to establish a three- and four-dimensional operating window.

For these reasons, a different approach will be proposed. We will explore this approach in one of the following sections. But first, before we can execute this approach, we still need to gather one additional piece of information.

## CAPABILITY OF CRITICAL INPUTS

We have now completed four of the five items we set out to complete. So far we have:

1. Identified the critical output variables
2. Developed measurements for these output variables and established specifications based on customer need;
3. Identified the critical input variables affecting the outputs;
4. Modeled the effect of the critical inputs on the critical outputs

Still to be accomplished is:

5. Determine manufacturing and supplier capabilities to control the critical input variables.

This requires performing capability studies and life testing on the different critical inputs. For hot bar temperature, a temperature transducer was installed to continuously record the temperature. The temperature range over several extended runs varied 6°F up and down from the set point. The range of 12°F is based on continuous measurement of the temperature giving thousands of readings and should conservatively contain

**Table 8** Manufacturing Capabilities of Critical Inputs

| Input | Estimated standard deviation |
|---|---|
| Hot bar temperature | 2°F |
| Dwell time | 0.08 sec |
| Pressure | 1 lb |
| Material temperature | 4°F |

at least 99.7% of future values or ±3 standard deviations. We will use one-sixth of this 12°F range, equal to 2°F, as a bound on the standard deviation.

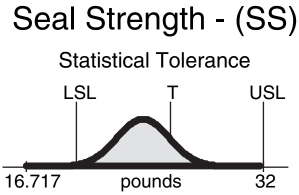
For DT, a high-speed video camera (100 frames per second) is used to measure the number of frames the hot bar is in contact with the material. This is performed repeatedly for several days. From this data, the DT standard deviation is estimated to be 0.08 seconds.

P was handled similar to HB. The estimated standard deviation is 1 lb. That leaves MT. The material sits in the room and is expected to vary by the same amount as RT. RT varies from summer to winter. Data would have to be collected over a year to determine the full range. As an alternative, operators were asked to remember the hottest and the coldest that the room becomes. This range was again divided by 6 to obtain an upper bound for the standard deviation equal to 4°F.

When obtaining estimates of the variation of the critical inputs, it is important to capture the full range of variation expected over long-term manufacturing. Table 8 summarizes the capabilities of the different critical inputs.

## ROBUST TOLERANCE ANALYSIS

We have now gathered all the necessary data. The next step is to use this information to develop the control plan. An approach called robust tolerance analysis (5) will be used. It is a combination of tolerance analysis for setting the width of the tolerances or operating windows of the inputs and robust design methodologies for setting the targets or nominals of the inputs. Robust tolerance analysis is a systematic approach for setting targets and tolerances to achieve a desired level of quality at



| Characteristic | Value |
|---|---|
| Average: | 24.206 |
| Standard Deviation: | 2.4963 |
| Cp: | 0.80 |
| Cc: | 0.30 |
| Cpk: | 0.56 |
| Def. Rate (normal) | 4.69% |

**Figure 5** Tolerance analysis when hot bar temperature = 175, dwell time = 0.5, pressure = 100, and materials temperature = 70. Abbreviations: LSL, lower spec limit; T, target; USL, upper spec limit.

the lowest possible cost. It starts with a tolerance analysis to evaluate the current design.

We have estimates of the variation for each of the inputs and a model for how the inputs affect the output. From this, we can predict how the output will vary. One way of doing this is to perform a computer simulation. This requires randomly generating values for the inputs and plugging them into the equation to see how the output behaves.

More generally, we can perform a tolerance analysis to predict how the output behaves. When all the inputs are targeted at the center of the study window, Figure 5 shows the results of a statistical tolerance analysis (HB = 175, DT = 0.5, $P$ = 100, and MT = 70). The seal strength is estimated to have an average of 24.2 lb with a standard deviation of 2.5 lb. The defect rate is estimated to be 4.69%.

This tolerance analysis was performed by taking the equation for the average from the response surface study and using it to derive the following equation for the seal strength standard deviation (6):

$$\sigma_{SS} = \sqrt{\begin{aligned}&(3.7514 - 0.0172t_{HB} - 0.9046t_{DT})^2\,\sigma_{HB}^2 + (263.30555 - 151.5974t_{DT} - 0.9046t_{HB})^2\,\sigma_{DT}^2\\ &+0.0408^2\,\sigma_P^2 + 0.0482^2\,\sigma_{MT}^2 + 0.00014792\sigma_{HB}^4 + 11490.885\sigma_{DT}^4 + 0.81830116\,\sigma_{DT}^2\end{aligned}}$$

The targets of the critical inputs are $t_{HB}$, $t_{DT}$, $t_P$, and $t_{MT}$ respectively, and these inputs vary around their targets with standard deviations $\sigma_{HB}$, $\sigma_{DT}$, $\sigma_P$, and $\sigma_{MT}$. This formula is based on the root sum of squares formula for combining variation. For each input, its contribution to the variation of seal strength is its slope times its standard deviation. The slopes can be obtained by taking partial derivatives of the equation for seal strength. The remaining terms are a result of the quadratic and interaction terms in the model.

Note that $\sigma_{SS}$ does not depend on $t_P$ and $t_{MT}$ but does depend on $t_{HB}$ and $t_{DT}$. Adjusting HB and DT affects the variation. Adjusting P and MT affects the average but not the variation. The capability study in Figure 5 was performed by entering the targets and standard deviations of the input variables from Table 8 into the equation for the average from the response surface study and the previous equation for the standard deviation.

In a previous section, an equation was fit to the log of the standard deviation using a response surface study. When HB = 175, DT = 0.5, P = 100, and MT = 70, the calculated standard deviation using this equation is 1.15. This is quite a bit lower than the standard deviation of 2.50 predicted using the tolerance analysis shown in Figure 5. What is the difference between these two estimates? The tolerance analysis estimate is based on the full range of variation expected in manufacturing. The response surface estimate is based on the variation observed during the limited study conditions. In general, the response surface estimate will underestimate the full range of variation. Taguchi (7) proposes a solution to this problem where for each trial, rather than take a random sample of units, a noise array is run where the inputs are purposely adjusted to mimic their variation under actual manufacturing conditions. This makes the study more difficult to run, especially if there are lots of sources of variation requiring a complex noise matrix. Robust tolerance analysis offers an alternative approach.

While the tolerance analysis estimate will generally be the larger estimate, this will not always be the case. The tolerance analysis estimate depends on having all the sources of variation included in the model. It, too, will underestimate the full range of variation in manufacturing if other sources of variation exist that were not included in the model. Best practice is to proceed with the approach that provides the larger estimate of the variation. In this case study, the predicted variation is larger than the observed variation, so we will proceed to use the tolerance analysis.

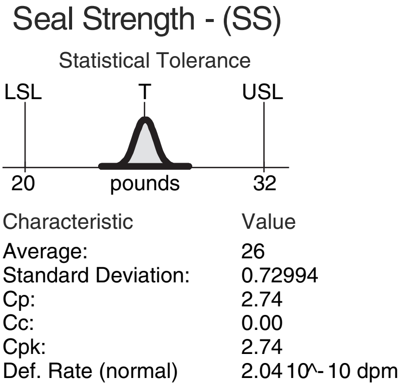
A tolerance analysis predicts the behavior of the output for a specified set of targets for the inputs. The tolerance analysis can be repeated for different sets of targets to identify the optimal targets. In this case study, we have an observed equation for the average, and we have a predicted equation for the standard deviation. These can be used to obtain an equation for the capability index $C_{pk}$. We can then maximize this equation to identify the optimal targets for the inputs.

The settings for the inputs maximizing $C_{pk}$ are shown in Table 9. The target for material temperature (MT) was fixed at 70 and the other three inputs optimized. This answers the question, without preheating the material, what are the optimal settings? The resulting performance is shown in Figure 6. The variation has been reduced by 71% from 2.50 to 0.73. The defect rate has been reduced from 4.7% to essentially zero. Adjusting targets on the input variables to reduce the variation of the output variables is called robust design. We have just seen an example of how robust designs can be achieved.

Maximizing $C_{pk}$ minimizes seal strength variation while centering the average at target. HB and DT were set to reduce the variation. P was set to center the average at 26 lb. Remember, the target of P does not affect the standard deviation. If seal strength shifts off-target, the operator should use

**Table 9** Optimal Targets for Critical Inputs

| Input | Target | Standard deviation |
|---|---|---|
| Hot bar temperature (HB) | 185°F | 2°F |
| Dwell time (DT) | 0.62 sec | 0.08 sec |
| Pressure (P) | 62 lb | 1 lb |
| Material temperature (MT) | 70°F | 4°F |



| Characteristic | Value |
|---|---|
| Average: | 26 |
| Standard Deviation: | 0.72994 |
| Cp: | 2.74 |
| Cc: | 0.00 |
| Cpk: | 2.74 |
| Def. Rate (normal) | 2.04 10^-10 dpm |

**Figure 6** Tolerance analysis for optimal conditions when material temperature = 70. *Abbreviations*: LSL, lower spec limit; T, target; USL, upper spec limit.

P to bring it back on target. The operators should not adjust HB or DT. Adjusting either will increase the variation.

Table 10 shows the three competing approaches to robust design. When the equation is known, a robust tolerance analysis should be performed. When the equation is not known, designed experiments must be used and all three

**Table 10** Different Approaches to Robust Design

| Name | How to estimate variation | Limitations |
|---|---|---|
| Dual response | Model standard deviation based on observed variation at each trial | Variation during limited study is frequently less than that in long-term manufacturing |
| Taguchi methods | Model standard deviation using noise array designed to mimic manufacturing variation | Underestimates variation if sources of variation are not included in noise matrix; works best if primary source(s) of variation have been identified |
| Robust tolerance analysis | Model average and uses a tolerance analysis to predict the variation | Underestimates variation if sources of variation are not included in the model |

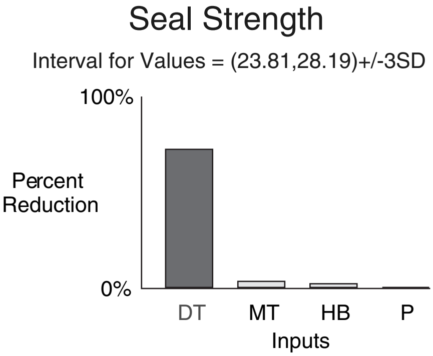


**Figure 7** Contributions of individual inputs. *Abbreviations*: DT, dwell time; HB, hot bar temperature; MT, material temperature; P, pressure.

approaches are applicable. The three approaches differ in how they obtain an equation for the standard deviation of the output. These approaches can be combined to generate even more efficient strategies. In this case study, a combination of the dual response and robust tolerance analysis was used. If the primary sources of variation have been identified, a combination of Taguchi methods and robust tolerance analysis might be preferred.

The optimal targets appear to result in excellent quality. This will not always be the case. If so, tighter tolerances must be specified for some of the critical inputs. To determine which inputs to tighten, the variation displayed in Figure 6 can be partitioned into the contribution of each input as shown in Figure 7. Tightening the tolerance of DT has the largest effect. A more detailed tolerance stack-up will be performed once a control plan is established. We will wait until then to determine if any tolerances need to be tightened.

## ESTABLISHING THE CONTROL PLAN

Based on the understanding gained, the control plan in Table 11 was established. The control plan provides operating ranges for each of the four critical inputs along with controls designed to ensure they remain within their operating windows. These

**Table 11** Initial Control Plan

| Name | Requirement | Controls |
|---|---|---|
| HB | Maintain between 179 and 191°F with a target of 185°F | Continuously monitor and alarm if out |
| DT | Maintain between 0.39 and 0.87 sec with a target of 0.63 sec | Continuously monitor and alarm if out |
| P | Maintain setting between 50 and 150 lbs; adjust in response to control chart of seal strength average<br>Once set, maintain within ±3 lb of set point | Continuously monitor and alarm if out |
| MT | Maintain between 60 and 80°F | Continuously monitor and alarm if out<br>Material must be in temperature-controlled environment at least 24 hours prior to use in production |
| Seal strength | Below 5% defective at worst-case conditions for critical inputs<br>Below 1% defective under normal conditions | Perform hourly inspections using variables sampling plan with AQL = 1.0% |
| Seal strength average | Average must be maintained between 26 ± 2 lb | Control chart using $\overline{X}$ at start-up and every hour using five samples. Adjust pressure as needed to keep in control limits |
| Seal strength standard deviation | Standard deviation must be below 2 lb ($C_p \geq 1$) at worst-case conditions for critical inputs<br>Standard deviation should be below 1 ($C_p \geq 2$) under normal conditions | Maintain S control chart in conjunction with $\overline{X}$ chart above to detect increase in process variation |

(*Continued*)

**Table 11** Initial Control Plan (*Continued*)

| Name | Requirement | Controls |
|---|---|---|
| Percent burns | Maintain below 5% defective at worst-case conditions for critical inputs<br>Below 1% defective under normal conditions | Perform hourly inspections using attribute sampling plan whose AQL = 1.0%. This will detect a major problem quickly<br>Trend inspection data on weekly basis to verify continual conformance to requirement |

*Abbreviations*: AQL, accepted quality level; DT, dwell time; HB, temperature hot bar; MT, material temperature; P, pressure.

operating windows are inside the operating window previously established for seal burn.

The widths of the operating windows were selected based on the capabilities of the four critical input variables combined with the optimal targets for seal strength. To confirm that these windows are adequate, a more detailed tolerance analysis will be performed. As ranges are specified for each of the inputs, and alarms are used to detect excursions, we cannot assume the different inputs remain centered around their targets. This is particularly true for RT where one could operate at either the low end or the high end of its range for considerable periods of time. This violates one of the assumptions of statistical tolerance, so a more appropriate tolerance analysis is required.

For HB, a target of 185°F and a tolerance of 179–191°F was specified based on historical data. Temperature cycles up and down. The average remains close to the target but not exactly equal to the target. As an alarm is used to ensure that the range is maintained, it was decided HB is best represented by the tolerance shown in Figure 8. It is expected that all values remain in the interval 179–191°F and that the average remains in the narrower interval 183–187°F. The interval for the average represents the middle one-third of the specification limits.

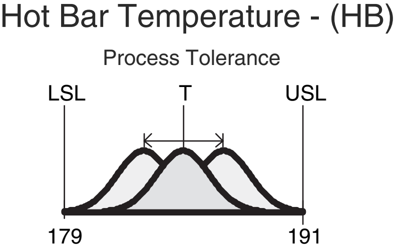


**Figure 8** Tolerance for hot bar temperature: values between 179 and 191°F and average in the middle one-third. *Abbreviations*: LSL, lower spec limit; T, target; USL, upper spec limit.

The type of tolerance shown in Figure 8 is referred to as a process tolerance (5). It is not as restrictive as a statistical tolerance, which assumes perfectly centered processes. It is more restrictive than a worst-case tolerance, which makes no assumptions about the average. Process tolerances like the one in Figure 8 represents middle ground between statistical tolerances and worst-case tolerances.

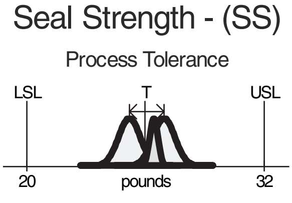
Likewise, process tolerances were established for DT and P. For P, there is not a specified target. The control plan uses a control chart of the seal strength average to detect seal strength being off-target and then allows P to be adjusted to maintain the average on target. For the tolerance stack-up, P will be set to the value in range of 50–150 lb which centers seal strength at 26 lb. For MT, a worst-case tolerance will be used. This makes no assumptions about MT other than it remains in the range of 60–80°F. A summary of the tolerances selected is shown in Table 12. The results of the tolerance stack-up are shown in Figure 9.

According to Figure 9, variation in the four critical inputs should not cause a problem if the seal strength is properly centered to start with. The variation in Figure 9 exceeds that in Figure 6, which was based on a statistical tolerance analysis. However, we are not yet done. Figure 9 assumes that the optimal set point for P is used. In actuality, there will be some error associated with selecting the set point for P.

**Table 12** Tolerances for Critical Inputs

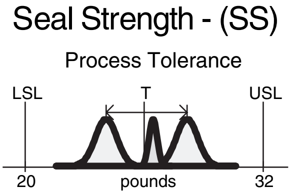
| Input | Type of tolerance | Target | Tolerance for values | Tolerance for average |
|---|---|---|---|---|
| Temperature hot bar | Process | 185°F | 179 to 191°F | 183 to 187°F |
| Dwell time | Process | 0.63 sec | 0.39 to 0.87 sec | 0.55 to 0.71 sec |
| Pressure | Process | Value which centers seal strength | Set point ± 3 lb | Set point ± 1 lb |
| Material temperature | Worst-case | 70°F | 60 to 80°F | |

The set point for P is set using an $\overline{\text{X}}$ control chart of seal strength based on a sample of five units. The sample size of five allows a shift of 1.5 standard deviations to be detected 70% of the time on the next point plotted. This should maintain the seal strength average within 1.5 standard deviations (=1.2 lb) of target. A worst-case tolerance of ±1.2 lb was added to the previous analysis to account for the adjustment error. The final results are shown in Figure 10. The worst $C_{\text{pk}}$ expected is 1.56 resulting in around two defects per million (dpm). This is a



| Characteristic | On-Target | Worst-Case |
|---|---|---|
| Average - Minimum: | 26.405 | 25.203 |
| Maximum: | | 26.928 |
| Standard Deviation: | 0.31373 | 0.82869 |
| Cp: | 6.37 | 2.41 |
| Cc: | 0.07 | 0.15 |
| Cpk: | 5.94 | 2.04 |
| Def. Rate (normal): | 1.92 10^-65 dpm | 0.000466 dpm |

**Figure 9** Tolerance analysis assuming optimal set point for pressure. *Abbreviations*: LSL, lower spec limit; T, target; USL, upper spec limit.



| Characteristic | On-Target | Worst-Case |
|---|---|---|
| Average -Minimum: | 26.405 | 24.033 |
| Maximum: | | 28.128 |
| Standard Deviation: | 0.31373 | 0.82869 |
| Cp: | 6.37 | 2.41 |
| Cc: | 0.07 | 0.35 |
| Cpk: | 5.94 | 1.56 |
| Def. Rate (normal): | 1.92 10^-65 dpm | 1.49 dpm |

**Figure 10** Tolerance analysis assuming seal strength centered using control chart. *Abbreviations*: LSL, lower spec limit; T, target; USL, upper spec limit.

level of quality called six-sigma quality. Under the normal operating conditions shown in Figure 10, the seal strength average is expected to be maintained within the window of 26 ± 2 lb and the standard deviation should remain below 1 lb. These requirements are listed in the control plan along with controls designed to ensure they continue to be meet.

The basic strategy used here is to establish operating windows for the critical input variables over which it is known that the seal strength standard deviation and seal burns will be acceptable. This operating window does not guarantee that the seal strength average is acceptable. Instead, a control chart is used with sufficient sample size to verify that the requirement for the seal strength average is met on a lot-by-lot basis.

## IQ TESTING AND OQ CHALLENGE TESTING

Having established the control plan, it is now time to perform the confirmation studies to demonstrate the control plan works, starting with installation qualification (IQ). During IQ, the alarms for HB, DT, P, and RT should be tested to demonstrate that they function properly.

Also, if not already done, the method used to measure seal strength should be validated. A control chart will be used to detect an off-center process. This control chart is required to detect a 2-lb shift off-target. It can do this if the seal strength standard deviation is below 1 lb. This requires that the measurement error for seal strength be below 0.5 lb.

Next is operational qualification (OQ) challenge testing. From Figure 4, the worst-case condition for seal burn occurs at the high settings of HB and DT which, per Table 12, are HB = 187°F and DT = 0.71 sec. A contour plot is shown in Figure 11 for the seal strength standard deviation over the selected operating windows for the average (targets). Worst-case conditions for seal strength standard deviation are HB = 187°F and DT = 0.71 seconds and HB = 183°F and DT = 0.55 seconds. Both these parameters simultaneously being at worst-case conditions is unlikely to occur in practice. Therefore, the defect level requirements will be relaxed somewhat at the challenge conditions.

For seal burn, it will be demonstrated with 90% confidence that the defect rate is below 5% defective at the worst-case condition. The double sampling plan for proportion nonconforming $n1 = 50$, $a1 = 0$, $r1 = 2$, $n2 = 75$, and $a2 = 2$ can be used for this purpose. This plan requires 50 samples initially. If there are no seal burns, it passes. If there are two or

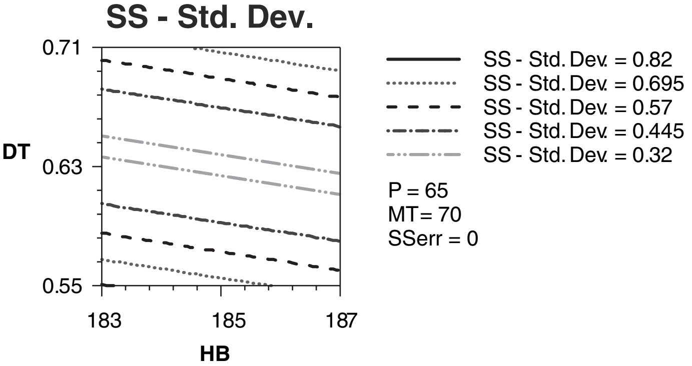


**Figure 11** Contour plot of seal strength standard deviation. *Abbreviations*: DT, dwell time; HB, hot bar temperature; SS, seal strength; MT, material temperature; P, pressure.

more seal burns, it fails. If a single seal burn is found, 75 additional units are inspected, and the total number of seal burns found must be two or less to pass. Passing this sampling plan provides 90% confidence that the seal burn percent defective is below 5% (8). If the defect rate is below 0.6%, there is a 95% chance of passing this plan. From Figure 4, the seal burn rate is expected to be below 1% at the worst-case conditions, so there is a high chance of passing.

For seal strength standard deviation, it will be demonstrated with 95% confidence that the standard deviation is below 2 lb ($C_p \geq 1$) at each of the worst-case conditions. The variables sampling plan for the standard deviation $n = 11$, $s = 1.4$ can be used for this purpose. This plan takes 11 samples and accepts if the estimated standard deviation is below 1.4. If this plans passes, one can state with 95% confidence that the standard deviation is below 2. If the standard deviation is below 1 lb, as expected, there is a 95% chance of passing this plan (9).

Also as part of the OQ worst-case challenges, it will be demonstrated that the control chart properly adjusts the average. At the two worst-case conditions for the standard deviation, once set-up adjustments are made and the initial point on the control chart is within the control limits, a variables sampling plan will be used to inspect for seal strength. Take 50 samples, calculate the capability index $P_{pk}$, and accept if the $P_{pk}$ is above 0.81. Passing this plan allows one to state with 95% confidence that the seal strength defect rate is below 2.5%. Again, a higher defect is allowed because the standard deviation is expected to be higher than under normal production conditions. If the defect rate is below 0.5%, there is a 95% chance of passing this plan.

The sampling plans are summarized in Table 13. Each plan was selected based on the claim that could be made if it passes. However, each plan was also evaluated with respect to what is required to assure the plan passes. Based on expected performance from previous studies, each plan has a reasonable chance of passing. It should be noted that the performance level must be significantly better than the performance level one is validating in order to have a reasonable chance of

**Table 13** Operational Qualification Challenge Testing

| Parameter tested | Claim | Sampling plan | 95% chance of passing if: |
|---|---|---|---|
| Seal burn | 90% confidence that the defect rate is below 5% | Double sampling plan for proportion nonconforming $n1=50$, $a1=0$, $r1=2$, $n2=75$, $a2=2$ | Defect rate is below 0.6% |
| Seal strength standard deviation | 95% confidence that the standard deviation is below 2 lb ($C_p \geq 1$) | Variables sampling plan for standard deviation $n=11$, $s=1.4$ | Standard deviation is below 1 lb |
| Seal strength out of spec | 95% confidence that the seal strength defect rate is below 2.5% | Variables sampling plan for proportion nonconforming $n = 50$, $P_{pk} \geq 0.81$ | Defect rate is below 0.5% |

passing. For seal burn, the sampling plan demonstrates that the rate of seal burns is below 5%. If the rate of seal burns is around 5% there is a 90% chance of failing. The seal burn rate must be below 0.6% (nine times better than 5%) to routinely pass the sampling plan.

## PQ TESTING

PQ testing involves three runs using materials purposely selected to represent the full range of materials used during manufacturing. The tests are summarized in Table 14. For percent burn, the double sampling plan $n1=250$, $a1=0$, $r1=3$, $n2=560$, and $a2=3$ will be used. Passing this sampling plan provides 90% confidence that the seal burn percent defective is below 1%. The samples are divided equally across the three PQ runs. This allows one to state that the process average across the lots is below 1%. If the seal burn percent defective is below 0.2%, there is a 95% chance of passing. Based on Figure 4, the seal burn rate should be well below this.

**Table 14** Performance Qualification Testing Under Anticipated Conditions

| Parameter tested | Claim | Sampling plan | 95% chance of passing if: |
|---|---|---|---|
| Seal burn | 90% confidence that the defect rate is below 1% | Double sampling plan for proportion nonconforming $n1 = 250$, $a1 = 0$, $r1 = 3$, $n2 = 560$, $a2 = 3$ | Defect rate is below 0.2% |
| Seal strength standard deviation | 95% confidence that the standard deviation is below 1 lb ($C_p \geq 2$) | Variables sampling plan for standard deviation $n=50$, $s=0.84$ | Standard deviation is below 0.7 lb |
| Seal strength out of spec | 95% confidence that the seal strength defect rate is below 1% | Variables sampling plan for proportion nonconforming $n=50$, $P_{pk} \geq 0.95$ | Defect rate is below 0.04% |

For seal strength standard deviation, the acceptance sampling plan used takes $n = 50$ samples and accepts if the estimated standard deviation is below 0.84. If this passes, one can state with 95% confidence that the standard deviation is below 1. If the standard deviation is below 0.7 lb, there is a 95% chance of passing this plan. According to Figure 10, we expect the standard deviation to be between above 0.3 but well below 0.7 lb. The sample size here is larger than needed but was matched to the sample size of the plan for the seal strength average next. This plan will be applied separately to each of the 3 PQ lots.

For seal strength average, take 50 samples and accept if the estimated $P_{pk}$ value is above 0.95. Passing this plan allows one to state with 95% confidence that the seal strength defect rate is below 1%. If the defect rate is below 0.04%, there is a 95% chance of passing. This plan will also be applied separately to each PQ lot.

Finally, it should be demonstrated that the four critical inputs variables remained within their limits. With good knowledge of the process performance, acceptance sampling plans can be selected that not only demonstrate the process is working, but for which there is a reasonable expectation of passing.

## CONCLUSION

Being able to validate a process first requires that a process be developed that "will consistently produce a product meeting its predetermined specifications and quality attributes." Validation, if fully integrated into the design process, is simply good business practice. Failure to properly execute these tools results in the design/validation process degenerating into one of trial and error. This is an unpredictable process with numerous surprises. Time is lost, the goal switches from obtaining the optimal design to just getting one that works, and the final stages of validation are approached with fingers crossed. Proper application of these tools results in the optimal design, reduces costs, improves quality, has few surprises, and decreases design time.

Designed experiments can play a key role in this process. They can be used to establish operating windows for attribute characteristics. However, their most important use is as part of the specification translation process. Designed experiments must be carefully integrated into the overall process. This process requires five items:

1. Identify the critical output variables.
2. Develop measurements for these output variables and establish specifications based on the customer need.
3. Identify the critical input variables affecting the outputs.
4. Model the effect of the critical inputs on the critical outputs.
5. Determine supplier and manufacturing capabilities to control the critical input variables.

Designed experiments play a key role in items 3 and 4. They also identify worst-case conditions for OQ testing and result in an understanding that allows more complex control plans to be established. However, designed experiments are not enough. They must be carefully coordinated with other tools like FMEA, mistake proofing, customer research tools, measurement system analysis, capability studies, acceptance sampling plans, tolerance analysis, robust design, and the numerous other tools described in Annex A of GHTF (2).

Based on the knowledge gained, a control plan can be established that ensures continual conformance of the process. All too often, validations are passed but, weeks later, major problems are encountered as materials change, tooling wears, operators change, and so on. These instances highlight that validation should not only demonstrate that things are currently OK, but should establish a system of controls designed to ensure ongoing conformance.

## REFERENCES

1. FDA. Guideline on General Principles of Process Validation. May 1987. Food and Drug Administration. (http://www.fda.gov/cder/guidance/pv.htm).
2. GHTF. Quality Management Systems—Process Validation Guidance. 2nd ed., January 2004. Global Harmonization Task Force (www.ghtf.org).
3. Taylor WA. Optimization and Variation Reduction in Quality. New York: McGraw Hill, 1991.
4. Box GEP, Draper NR. Empirical Model Building and Response Surfaces. New York: John Wiley & Sons, 1987.
5. Taylor WA. VarTran User Manual. Libertyville, IL: Taylor Enterprises, 2004 (http://www.variation.com/vta).
6. Cox ND. How to Perform Statistical Tolerance Analysis. Milwaukee: ASQC Quality Press, 1986.
7. Taguchi G. System of Experimental Design. Dearborn, Michigan: American Supplier Institute, 1987.
8. Taylor WA. Guide to Acceptance Sampling. Libertyville, IL: Taylor Enterprises, 1992.
9. Taylor WA. Sampling Plan Analyzer User Manual. Libertyville, IL: Taylor Enterprises, 2005 (http://www.variation.com/spa).

8

# Efficient and Effective Process Development Work: The Key to a Successful Validation Study

**JEFFREY T. FIELD**

J. T. Field Consulting Services, LLC
Woodbury, Connecticut, U.S.A.

## INTRODUCTION

Reducing the amount of time required to properly develop, scale-up, and optimize a new process and/or product is absolutely critical and directly related to the competitive position of any business. Thus, it is important that all product development work be performed in a highly efficient and cost-effective manner. Initial product development work must identify the critical inputs to a manufacturing process and set proper specifications for these variables. In addition, if the process is shown to be robust, the tolerances on noncritical variables can be widened.

In any industry regulated by the Food and Drug Administration (FDA), it has historically been very difficult to make changes to a product's formulation, specifications, and/or processing conditions once they have already been established and documented in previous development or clinical work. Consequently, there tends to be a "snowballing" effect where poor or inadequate initial development work results in a final commercial product that is out of control and often fails subsequent process validation studies. By using "Design of Experiments (DOE)" from the beginning of process development to the end, the robustness of the manufacturing process can be maximized while minimizing the time required to bring the product to market. In addition, by using DOE upfront to gain a thorough understanding of the process, the scope of the required validation work can be easily identified and justified. This leads to efficient process validation studies with a high probability of success.

This book is dedicated to showing the reader how to use DOE during validation studies. However, the additional purpose of this particular chapter is to show the awesome power and value of DOE when used prior to validation studies. In other words, DOE is the best tool to use during initial process development in order to effectively and efficiently characterize the process and maximize the chances for success during future validation runs and commercial manufacturing of the product. The basic idea is to use DOE to gain a solid understanding of the key factors that affect a certain response variable. For example, how is a tablet's dissolution rate (response variable) affected by different factors, such as active-ingredient particle size, blend time, excipients, granulation method, and the like? At the beginning of any process development phase, it is not uncommon to have a list of 5, 10, or even 20 potential factors that could affect a critical response variable. The proper use of designed experiments is the fastest and most effective way to whittle this list down to only the factors that actually have an impact on the response. This results in much faster and cost-effective validation studies because only the critical factors are included in the final validation work. The previously written "Development Report" will contain the DOE work which shows that the other factors do not affect the response variable over the ranges studied. This upfront elimination of potential factors is very important because it makes the final validation work more clear and concise. Of course, having fewer factors to consider when writing the validation protocol makes the final workload much more reasonable and not so time consuming.

As it is often the best and easiest to learn from examples, this chapter contains two different case studies to illustrate the proper use and benefits of statistically designed experiments during process development work (performed prior to actual validation studies). In order to provide a template for performing a good DOE study, each case study will follow the same general sequence of steps as follows:

*Step 1*: Make a list of all factors that could affect the desired response variable(s). This is usually performed in a meeting room using a brainstorming process based on past experience with the process, previous data, and/or common sense.

*Step 2*: Choose the "High" and "Low" levels for each factor. For example, if pressure is one of the factors, it could have a "High" value of 40 psi and a "Low" value of 20 psi. It is alright to have a factor that is an attribute too. For example, if the factor is the material used, then the "High" level could be Material A and the "Low" level could be Material B (or vice-versa; in this case, the assignment of "High" and "Low" levels is arbitrary).

*Step 3*: Choose the design for your experiment. Fractional factorial designs or low-resolution designs are best for process development work where there are several (say four or more) factors to consider. Full factorial designs are used when it is necessary to eliminate all confounding or aliases between main effects and interactions.

*Step 4*: Put together the design matrix. In the case studies that follow, MINITAB™ software will be used for the matrix design and data analysis. MINITAB will automatically put the experimental runs in random order. This is important because randomization will help to eliminate any biases caused by running the same conditions several times in a row.

*Step 5*: Determine the sample size for each individual experimental run.

*Step 6*: Perform the experiments and compile the data. Wherever possible, use a good software package to analyze the data (e.g., MINITAB).

*Step 7*: Use the data to draw conclusions.

*Step 8*: If necessary, add more runs to the design matrix to eliminate aliases or confounding patterns. For example, if a fractional factorial experiment shows evidence of interactions between variables, it may be necessary to run the full factorial to determine which interactions are truly important.

*Step 9*: Perform confirmation runs and/or replicate the experiment. This is an extremely important part of the DOE process. As the scope of the experimentation work

> is usually limited by economic as well as other factors, this step is, unfortunately, often times overlooked. There are many different ways of doing confirmation experimentation and it will always depend on the given situation. A good deal of thinking is required at this step. Ultimately, you need to assure yourself that you have good data and a good understanding of your process. If you do not have this, then all the original work is of little value.

Confirmation runs often take the form of a replication (one or more times) of the original matrix. This means running the same experiments over again in a new random order and collecting the response data again. Or, depending on the situation, one might just do some experimental runs using the best factor settings as predicted by the experimental results from the original matrix. In the case of matrix replication, it is important to note that there is a difference between rerunning or replicating the experiments versus just repeating measurements during the initial experimental runs. This is because several measurements from the same experimental run would tend to have a smaller standard deviation compared with the data from several separate runs. Thus, better experimental data is usually obtained by replicating the design matrix (as opposed to just using repeated measurements).

Depending on the situation, confirmation runs may be done prior to validation runs. Or, if you're feeling really good about the data and your understanding of the process, the validation runs (per the written protocol) may be the confirmation runs for the DOE. Remember, validation should always be just a confirmation of what you already know. Thus, you should almost never fail a validation study. Right?

Before we begin, a general disclaimer must be made. As is the case with life in general, DOE offers no guarantees for success. It is important to realize that there is no magical system or methodology out there that will solve all of our problems every time. Although DOE is the best way to perform process development work, there is absolutely never a time when an experimenter can stop using his or her brain. One must never stop thinking. If a given data set does not, from a logical standpoint, seem to make sense, then there is a good chance that there is a problem with the experiment. In addition, one must always be aware of the difference between a data set's statistical significance versus practical significance. The bottom line is, use DOE as a powerful tool, but don't ever stop thinking. That being said, let's have some fun with some real case studies now.

## Case Study 1—Elimination of Leakage Between Welded Parts

The following is an excellent example of a typical designed experiment performed early in the development process (i.e., long before process validation studies). It is based on an actual study performed at a medical device manufacturing plant.

This example includes a list of eight potential factors that might affect the final product characteristic or response variable. Without the use of DOE, gaining an understanding of how eight different factors come together to affect the final product would be a truly daunting task at best. However, the process of determining which factors actually have an impact on the response variable may be greatly simplified by applying a fractional factorial DOE.

The process under study is a typical process that one might see at a medical device manufacturing site. In this case, it is a simple ultrasonic welding process where two plastic parts are being welded together to form a strong bond. The response variable to be considered is the number of parts that have a crack in the weld area after processing. So let's go through our list of steps.

*Step 1*: List the factors and response variable(s).

There are eight factors related to the set-up of the welding station that might cause cracks in the final welded part. In order to keep it simple, we will just label factor A through factor H in our design matrix to follow. It is important to note that this process actually has more than one response variable. In fact, that is usually the case during process development and validation work. For example, the strength of the weld and the number of parts that are discolored could be other important responses. However, in order to keep things simple, we will only consider the one response for this particular case study. That is, the number of parts with cracks when observed visually using a magnifying glass.

*Step 2*: Determine the factor levels.

Each factor is assigned a "High" and "Low" level. For this particular case study, in order to show that the actual factors and levels do not affect the final data analysis, we'll leave the factors in a coded format in the design matrix. A high level for a given factor is indicated by a numeral "1" and a low level is indicated by a "–1". See Table 1.

*Step 3*: Choose the design matrix.

We will use Figure 1 to choose the matrix for this designed experiment. The columns are labeled with the number of factors, and the rows are labeled with the number of experimental runs that will be required to complete the DOE. In each individual box, the design resolution is given. A detailed discussion of design resolution is beyond the scope of this book. However, the boxes are labeled to assist in the choosing process. For a typical process development study, it is usually best to start off with a design having resolution IV. In this type of design, the calculated main effects will be confounded (or aliased) only with three-factor interactions. As the existence of three-factor interactions in nature is

No. of Factors

| No. of Runs | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 4 | Full | III | | | | | | | |
| 8 | | Full | IV | III | III | III | | | |
| 16 | | | Full | V | IV | IV | IV | III | III |
| 32 | | | | Full | VI | IV | IV | IV | IV |
| 64 | | | | | Full | VII | V | IV | IV |
| 128 | | | | | | Full | VIII | VI | V |

**Figure 1** Available factorial designs (with resolution).

**Table 1** Design Matrix for Case Study 1

| Std. order | Run order | Factor A | Factor B | Factor C | Factor D | Factor E | Factor F | Factor G | Factor H | Crack rate |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 15 | –1 | –1 | –1 | –1 | –1 | –1 | –1 | –1 | |
| 2 | 2 | 1 | –1 | –1 | –1 | –1 | 1 | 1 | 1 | |
| 3 | 12 | –1 | 1 | –1 | –1 | 1 | –1 | 1 | 1 | |
| 4 | 18 | 1 | 1 | –1 | –1 | 1 | 1 | –1 | –1 | |
| 5 | 10 | –1 | –1 | 1 | –1 | 1 | 1 | 1 | –1 | |
| 6 | 11 | 1 | –1 | 1 | –1 | 1 | –1 | –1 | 1 | |
| 7 | 6 | –1 | 1 | 1 | –1 | –1 | 1 | –1 | 1 | |
| 8 | 1 | 1 | 1 | 1 | –1 | –1 | –1 | 1 | –1 | |
| 9 | 16 | –1 | –1 | –1 | 1 | 1 | 1 | –1 | 1 | |
| 10 | 4 | 1 | –1 | –1 | 1 | 1 | –1 | 1 | –1 | |
| 11 | 13 | –1 | 1 | –1 | 1 | –1 | 1 | 1 | –1 | |
| 12 | 7 | 1 | 1 | –1 | 1 | –1 | –1 | –1 | 1 | |
| 13 | 9 | –1 | –1 | 1 | 1 | –1 | –1 | 1 | 1 | |
| 14 | 19 | 1 | –1 | 1 | 1 | –1 | 1 | –1 | –1 | |
| 15 | 3 | –1 | 1 | 1 | 1 | 1 | –1 | –1 | –1 | |
| 16 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 17 | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 18 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 19 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |

*Note*: 1, “High” level; –1, “Low” level; 0, center point.

extremely rare, it is often safe to discount the three-factor interaction if it does not make sense from a logical standpoint. The potential problem with these designs is that the two-factor interactions are confounded with each other. So, if a significant two-factor interaction is found, the experimenter will not know specifically which one is important. This is the price one must pay for doing less than a full-factorial design. All full-factorial designs and other designs having a resolution of at least V are labeled as such in Figure 1. This means that they have minimal or no confounding of effects. For a late-stage process development study having just a few factors or a validation study that looks to establish "proven acceptable ranges" for the different factors, a full factorial design is often the best choice because all possible combinations of factors are studied and none of the calculated effects are confounded with other effects. However, it is not usually the best choice for screening experiments because they require at least twice as many experimental runs and, thus, much more time and effort than a fractional factorial design. Also, if there are no interactions present, then the resolution IV design will likely provide the same information, but with far less experimental runs (at least half as many). Designs having resolution III are normally not a good choice because the main effects are confounded with potential two-factor interactions that may or may not exist.

Looking at Figure 1, as the number of factors in this particular case study is eight, we will choose the first resolution IV design that we encounter when moving down the column. This design requires 16 experimental runs. This is known as a $2^{8-4}$ fractional factorial design. A $2^8$ full-factorial design would require 256 runs. Thus, the $2^{8-4}$ fractional factorial is one sixteenth of a full design.

*Step 4*: Set up the design matrix.

MINITAB is used to set up the design matrix. The matrix showing the 16 experiments to be run is shown in Table 1. The column labeled "Run Order" is used to randomize the experiments to eliminate any potential bias caused by the order in which the experiments are run. In order to calculate an error estimate within the data, three center-point runs are included within the matrix. Thus, the actual number of runs needed is 19. Please note that a better way to get a good error estimate is to replicate the experimental runs at least two or more times. However, this is not always possible from a practicality standpoint. Also, note that the last column (labeled "Crack Rate") in Table 1 is reserved for the actual response data for each run.

*Step 5*: Determine the sample size for each run.

Note that our response "variable" is really not a variable at all, but an attribute. Each individual part is either good or bad. In other words, each part either has a crack or does not have a crack. We are not getting any numerical data from each part. Thus, we are confronted with a dilemma that has long plagued would-be experimenters. How do we perform a statistically designed experiment with attribute data? The answer is somewhat simple although often difficult to deal with: we must change the attribute data into variable data. A key assumption in the analysis of factorial designs is that the response is measured on a continuous scale and has a constant variance. Since attribute data is not measured on a continuous scale and does not have a constant variance, the data must be transformed to meet these requirements. In this case, our attribute data will be changed to a continuous variable by first calculating a fraction defective (or "Crack Rate") and then applying a variance stabilizing transformation. Therefore, we will divide, for each run, the number of defective parts (i.e., parts with observed cracks) by the total number of parts inspected. In this way, we are calculating a "fraction defective" (p) that can be used in our subsequent data analysis. Each fraction defective will then be transformed using an arc-sine root transformation[1] given by the following formula:

$$R = \sin^{-1} \sqrt{p} \text{ (or } R = \text{Arc-Sine}(p^{1/2}))$$

where:
R = the final Response Variable to be analyzed
p = the calculated fraction defective

By applying the arc-sine root transformation to the fraction defective, the final response data to be analyzed is more likely to comply with the assumptions of traditional DOE.

Often times, the amount of attribute data needed to calculate a statistically significant proportion (e.g., fraction defective) is rather large. In this case, the sample size will be determined using MINITAB software. Assuming we are interested in detecting a difference between a sample that is 4% defective and a sample that is 1% defective (note: this is a decision that the experimenter and/or company management must make), the MINITAB “Power and Sample Size” calculator for two proportions can be used. This results in a sample size of 568 parts per run that will have to be inspected for cracks. Fortunately, for this particular case, the visual inspection of each part takes only a few seconds to complete. So the inspection of 568 parts actually takes less than one hour.

*Step 6*: Perform the experiments/compile the data.

After running each of the 19 experiments given in Table 1, the data is added to the design matrix. Columns are added to the MINITAB worksheet for the calculated fraction defective and the final transformed data. The transformed data (using the arc-sine root transformation) is then analyzed using MINITAB. The actual response variable data is shown in Table 2. The transformed data is in the column labeled “ArcSine”.

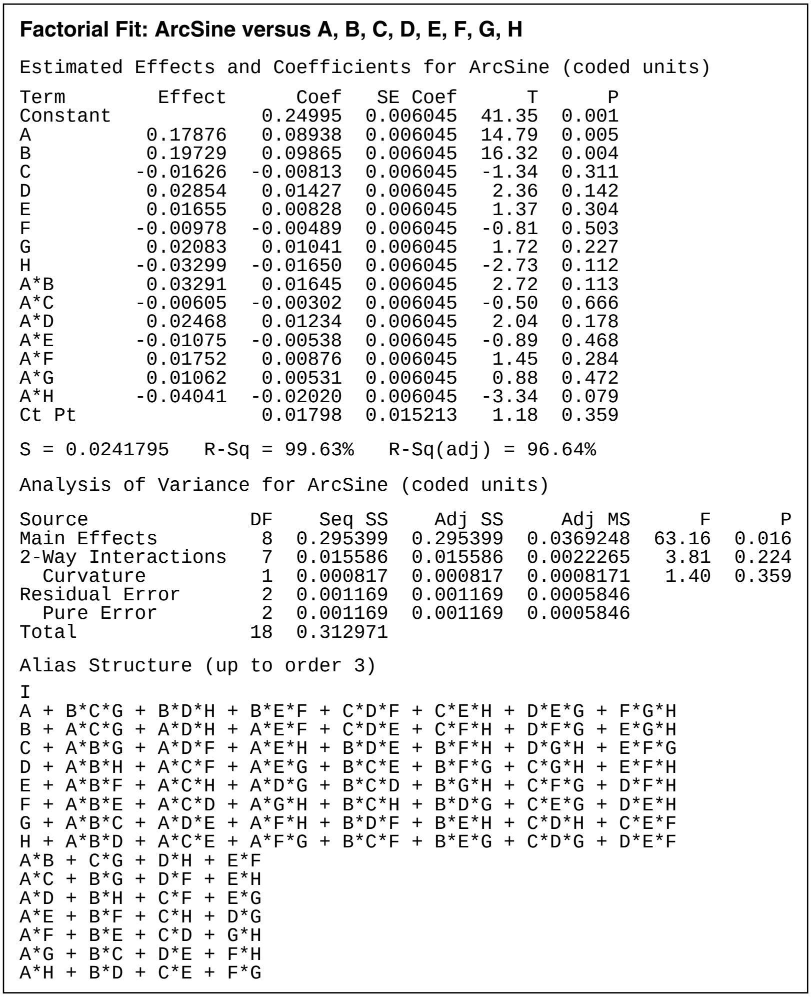
*Step 7*: Analyze the data/draw conclusions.

The resulting print-out from MINITAB is given in Figure 2. However, the easiest way to interpret the data is to look at the pareto chart of calculated effects (Fig. 3) and related normal probability plot (Fig. 4). The vertical line on the pareto chart in Figure 3 corresponds to a $P$-value of 0.10 for each calculated effect. In other words,

**Table 2** Design Matrix for Case Study 1 with Response Data

| Std.<br>order | Run<br>order | Factor<br>A | Factor<br>B | Factor<br>C | Factor<br>D | Factor<br>E | Factor<br>F | Factor<br>G | Factor<br>H | Crack<br>rate | ArcSine |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 15 | –1 | –1 | –1 | –1 | –1 | –1 | –1 | –1 | 0.0052 | 0.0727 |
| 2 | 2 | 1 | –1 | –1 | –1 | –1 | 1 | 1 | 1 | 0.0352 | 0.1887 |
| 3 | 12 | –1 | 1 | –1 | –1 | 1 | –1 | 1 | 1 | 0.0774 | 0.2820 |
| 4 | 18 | 1 | 1 | –1 | –1 | 1 | 1 | –1 | –1 | 0.2024 | 0.4667 |
| 5 | 10 | –1 | –1 | 1 | –1 | 1 | 1 | 1 | –1 | 0.0052 | 0.0727 |
| 6 | 11 | 1 | –1 | 1 | –1 | 1 | –1 | –1 | 1 | 0.0176 | 0.1330 |
| 7 | 6 | –1 | 1 | 1 | –1 | –1 | 1 | –1 | 1 | 0.0422 | 0.2070 |
| 8 | 1 | 1 | 1 | 1 | –1 | –1 | –1 | 1 | –1 | 0.1989 | 0.4623 |
| 9 | 16 | –1 | –1 | –1 | 1 | 1 | 1 | –1 | 1 | 0.0070 | 0.0840 |
| 10 | 4 | 1 | –1 | –1 | 1 | 1 | –1 | 1 | –1 | 0.0950 | 0.3134 |
| 11 | 13 | –1 | 1 | –1 | 1 | –1 | 1 | 1 | –1 | 0.0492 | 0.2238 |
| 12 | 7 | 1 | 1 | –1 | 1 | –1 | –1 | –1 | 1 | 0.1760 | 0.4329 |
| 13 | 9 | –1 | –1 | 1 | 1 | –1 | –1 | 1 | 1 | 0.0070 | 0.0840 |
| 14 | 19 | 1 | –1 | 1 | 1 | –1 | 1 | –1 | –1 | 0.0669 | 0.2616 |
| 15 | 3 | –1 | 1 | 1 | 1 | 1 | –1 | –1 | –1 | 0.0651 | 0.2580 |
| 16 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.1936 | 0.4556 |
| 17 | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0704 | 0.2685 |
| 18 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0580 | 0.2434 |
| 19 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0827 | 0.2917 |

*Note*: 1, “High” level; –1, “Low” level; 0, center point.



**Factorial Fit: ArcSine versus A, B, C, D, E, F, G, H**

```
Estimated Effects and Coefficients for ArcSine (coded units)

Term         Effect      Coef   SE Coef       T      P
Constant              0.24995  0.006045   41.35  0.001
A           0.17876   0.08938  0.006045   14.79  0.005
B           0.19729   0.09865  0.006045   16.32  0.004
C          -0.01626  -0.00813  0.006045   -1.34  0.311
D           0.02854   0.01427  0.006045    2.36  0.142
E           0.01655   0.00828  0.006045    1.37  0.304
F          -0.00978  -0.00489  0.006045   -0.81  0.503
G           0.02083   0.01041  0.006045    1.72  0.227
H          -0.03299  -0.01650  0.006045   -2.73  0.112
A*B         0.03291   0.01645  0.006045    2.72  0.113
A*C        -0.00605  -0.00302  0.006045   -0.50  0.666
A*D         0.02468   0.01234  0.006045    2.04  0.178
A*E        -0.01075  -0.00538  0.006045   -0.89  0.468
A*F         0.01752   0.00876  0.006045    1.45  0.284
A*G         0.01062   0.00531  0.006045    0.88  0.472
A*H        -0.04041  -0.02020  0.006045   -3.34  0.079
Ct Pt                 0.01798  0.015213    1.18  0.359

S = 0.0241795   R-Sq = 99.63%   R-Sq(adj) = 96.64%

Analysis of Variance for ArcSine (coded units)

Source              DF    Seq SS    Adj SS     Adj MS      F      P
Main Effects         8  0.295399  0.295399  0.0369248  63.16  0.016
2-Way Interactions   7  0.015586  0.015586  0.0022265   3.81  0.224
  Curvature          1  0.000817  0.000817  0.0008171   1.40  0.359
Residual Error       2  0.001169  0.001169  0.0005846
  Pure Error         2  0.001169  0.001169  0.0005846
Total               18  0.312971

Alias Structure (up to order 3)

I
A + B*C*G + B*D*H + B*E*F + C*D*F + C*E*H + D*E*G + F*G*H
B + A*C*G + A*D*H + A*E*F + C*D*E + C*F*H + D*F*G + E*G*H
C + A*B*G + A*D*F + A*E*H + B*D*E + B*F*H + D*G*H + E*F*G
D + A*B*H + A*C*F + A*E*G + B*C*E + B*F*G + C*G*H + E*F*H
E + A*B*F + A*C*H + A*D*G + B*C*D + B*G*H + C*F*G + D*F*H
F + A*B*E + A*C*D + A*G*H + B*C*H + B*D*G + C*E*G + D*E*H
G + A*B*C + A*D*E + A*F*H + B*D*F + B*E*H + C*D*H + C*E*F
H + A*B*D + A*C*E + A*F*G + B*C*F + B*E*G + C*D*G + D*E*F
A*B + C*G + D*H + E*F
A*C + B*G + D*F + E*H
A*D + B*H + C*F + E*G
A*E + B*F + C*H + D*G
A*F + B*E + C*D + G*H
A*G + B*C + D*E + F*H
A*H + B*D + C*E + F*G
```

**Figure 2** MINITAB™ print-out for case study 1. *Abbreviations*: DF, degrees of freedom; SS, sum of squares; MS, mean square.

if there is at least a 90% probability that a factor is statistically significant in affecting the response variable (“ArcSine”), then the bar for that particular factor will extend beyond the red line. These main effects and/or interaction terms are the only points that are labeled by MINITAB on the normal probability plot in Figure 4.

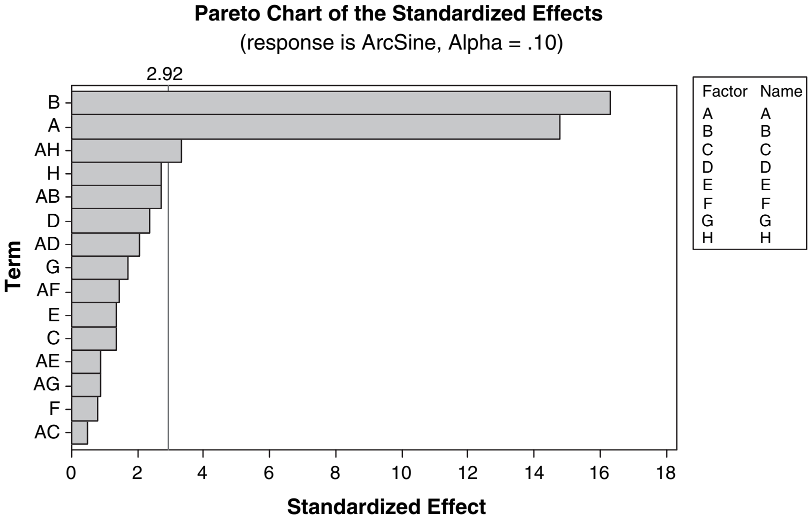


**Figure 3** Pareto chart for case study 1.

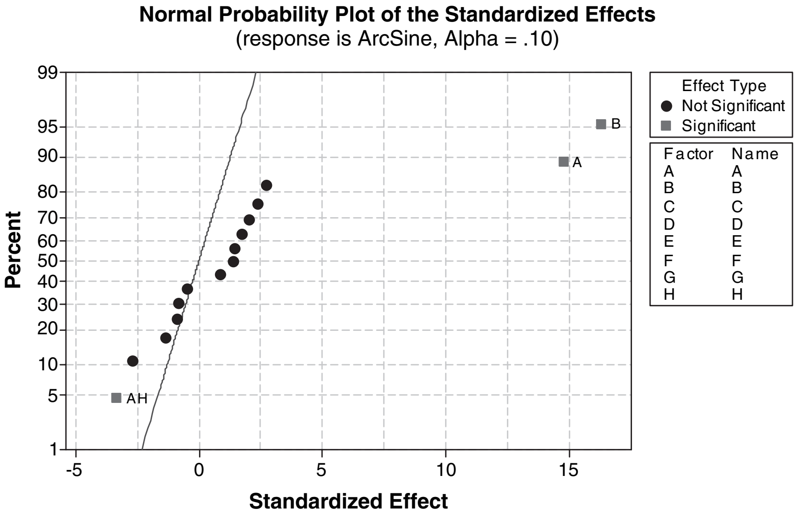


**Figure 4** Normal probability plot for case study 1.

> Together, these plots show that factors A and B are by far the most significant in affecting the response variable. This same conclusion may also be drawn from the "Main Effects Plots" shown in Figures 5 and 6 where the most significant factors have a plot with a large slope. In addition, there is an apparently significant interaction between factors A and H (Figs. 3, 4, and 7). Statistically, the data is not saying that factors H or the A–H interaction is definitely important. However, as we are preparing for future validation studies and we do not want to overlook anything, we will take a more conservative route and include factor H in future work.

So, in summary, the final conclusions are:

1. Factors A and B are, by far, the most important in affecting the proportion of parts that end up with cracks. The lower values for these factors result in less cracks in the parts. The main effects plots for all factors are given in Figure 5. The Main Effects plots for only the statistically significant factors

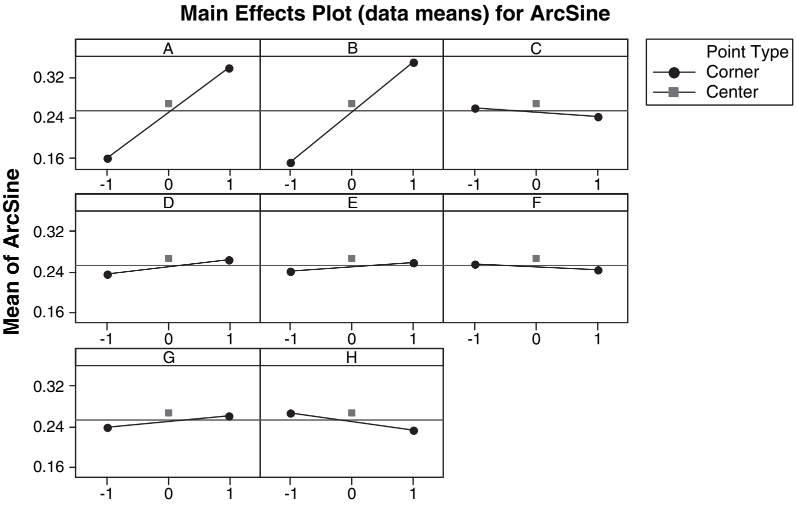


**Figure 5** Main effect plots for all factors (A–H).

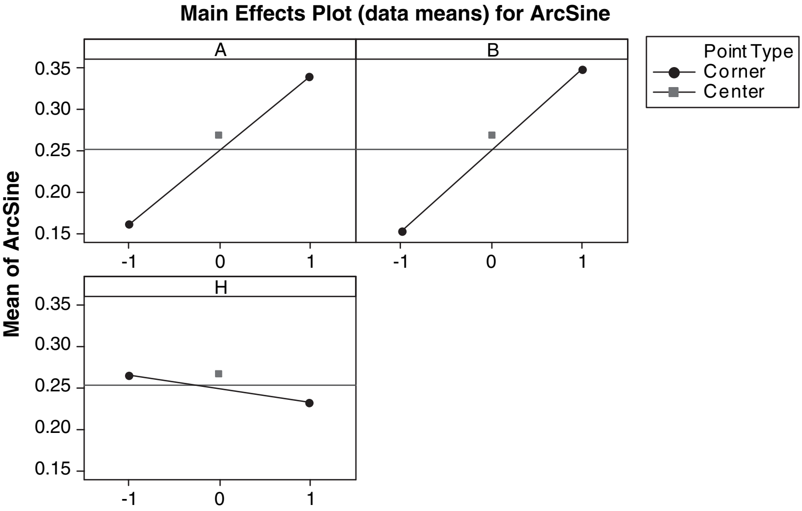


**Figure 6** Main effect plots for statistically significant effects (A, B, and H).

(i.e., Factors A, B, and H) are given in Figure 6. Note that the Main Effects plots for factors A, B, and H are not horizontal. This indicates statistical significance.

2. There is potentially a minor interaction between factors A and H. This is illustrated by the plot given in Figure 7. If there were no interaction, the two lines would be essentially parallel. Since the existence of this interaction makes sense from a logical perspective, Factor H will not be discounted from future studies.
3. All of the other factors (at their given high and low levels) do not appear to have a significant impact on the response variable. It may make sense to do further experiments to try to widen the tolerance on some or all of these factors. This will make the process more robust in the end. This may also allow these factors to be set at their most economic levels during commercial manufacturing.

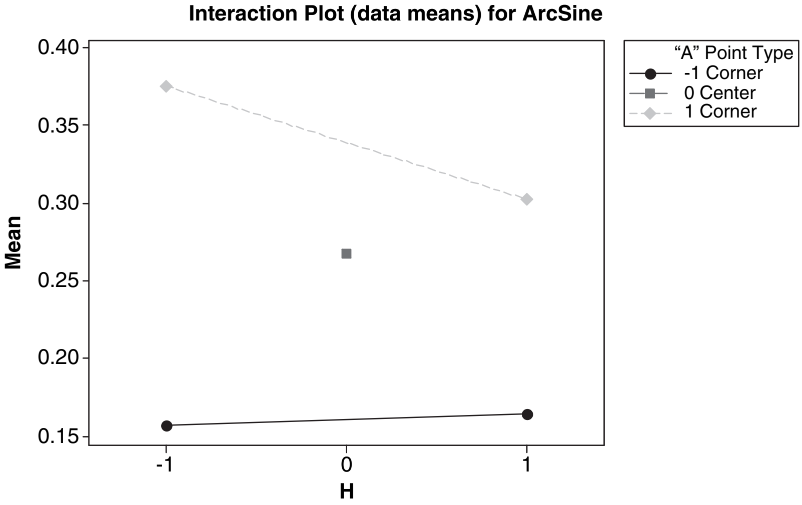


**Figure 7** Interaction plot for factors A and H.

*Step 8*: Eliminate or reduce aliases.

The "alias structure" for the DOE is given by MINITAB in Figure 2. This shows that the A–H interaction is actually the cumulative effect of the A–H component along with the B–D, C–E, and F–G interactions. In other words, the A–H interaction is confounded with the B–D, C–E, and F–G interactions. Remember that this is the price we paid for doing less than the full factorial design. At this point, it may make sense to think about the different factors logically and decide what interactions actually make sense. Which of these two-factor interactions are highly unlikely? For simplicity sake, we will assume that only the A–H interaction makes any sense for this particular example. This is somewhat normal in that, when analyzing a typical designed experiment, it is often the case that the most significant individual factors have a significant interaction term as well. However, if there was any doubt, further experimentation would be necessary to remove the aliases. This is what makes DOE so great. The experiments are like building blocks where you start with the minimum amount of work needed. Then, if necessary, you can always add more experiments later. In this example, if we wanted to make sure that no two-factor interactions were confounded with other two-factor interactions, we would have to move up to a Resolution V design. According to Figure 1, we would have to add 48 more runs to the original design matrix (64 total runs). Again, in order to keep things simple, we will assume that we are happy with the data "as is" and move on to confirmation runs.

*Step 9*: Confirm the results.

When deemed appropriate for an early process development study, the confirmation of the original data and conclusions can be obtained by replicating the original design matrix. The confirmation of the quantitative impact of factors A, B, and H may be obtained by running a new DOE using only these three variables while holding the others constant at their target values. Again using Figure 1, the resultant DOE matrix would be a full factorial ($2^3$) requiring eight runs.

Finally, the validation work may be performed for this process. This would most likely consist of repeating the $2^3$ full-factorial design mentioned earlier along with three process validation runs at nominal conditions for all factors. The process validation runs would essentially be confirmation runs of the actual process to be used in future commercial manufacturing. Of course, all of these experiments would be based on an approved protocol and would have to yield acceptable data compared with the product specifications.

Whereas case study 1 is a good example of work performed early in the product development life cycle, the next case study presented in this chapter is more representative of a study performed much later during the development stage (or even during process/product validation). From a DOE perspective, the only real difference between the two case studies is the number of factors being considered. As case study 1 had eight factors, this next example has only three factors to examine. This is typical for late-stage process development and validation work because the list of potential factors has already been whittled down from the "trivial many" to the "critical few."

### Case Study 2—Process Development and Validation of an Adhesive Dispensing Station

Like the previous example, this case study is based on an actual designed experiment performed at a medical device manufacturing plant. The process under scrutiny should be somewhat simple to visualize. Two separate parts of a medical device are to be essentially "glued" together using a liquid adhesive. In this case, the two parts are a needle and a plastic housing used to hold the needle. The "adhesive dispensing station" is just one station on an assembly line that makes the entire device. The two parts arrive at the station where the adhesive is applied to permanently attach the needle to the housing. Given adequate time for the adhesive to dry, the response variable to be considered is the bond strength or the force required to pull the needle from the housing by a special testing instrument. There are three factors to be considered in the DOE.

Again, we will go through our list of steps…

*Step 1*: List the factors and response variable(s).

There are three factors related to the set-up of the adhesive dispensing station that may impact the final bond strength. They are as follows:

1. Time = the time (in seconds) that the adhesive is allowed to flow from an application nozzle onto the part.
2. Pressure = the level of pressure used to dispense the adhesive.
3. Vacuum Level = the level of vacuum used to ensure proper application of the adhesive.

*Step 2*: Determine the factor levels.

Each factor is assigned a "High" and "Low" level. For this example, the coded format will not be used. Instead,

**Table 3** Design Matrix for Case Study 2

| Standard order | Run order | Time (sec) | Pressure (psi) | Vacuum level | Bond strength (lbs) |
|---|---|---|---|---|---|
| 1 | 7 | 0.2 | 10 | 5 | |
| 2 | 3 | 0.4 | 10 | 5 | |
| 3 | 9 | 0.2 | 30 | 5 | |
| 4 | 2 | 0.4 | 30 | 5 | |
| 5 | 8 | 0.2 | 10 | 15 | |
| 6 | 4 | 0.4 | 10 | 15 | |
| 7 | 1 | 0.2 | 30 | 15 | |
| 8 | 6 | 0.4 | 30 | 15 | |
| 9 | 5 | 0.3 | 20 | 10 | |

the actual high and low values for each factor are listed in Table 3.

*Step 3*: Choose the design matrix.

Again, we will use Figure 1 to choose the matrix for this designed experiment. As the number of factors in this case study is three, we really have no choice but to use the full-factorial design having eight runs. The half-fraction design having four runs is a resolution III design and, therefore, of little value at this stage.

*Step 4*: Set up the design matrix.

MINITAB is used to set up the design matrix. The matrix showing the experiments to be run is shown in Table 3.

The last column is reserved for the actual data for each run. Note that a center-point run (run 9) is included in the matrix. So the number of experiments is nine (i.e., the eight experiments required for the full-factorial design plus the one center-point run). The purpose of this center-point run is to gain an understanding of the possible curvature of the response variable data. In other words, we want to know if the bond strength is linear over the full range of factors studied.

At this point, we need to consider how we will separate actual "effects" from random error within the experiment. In case study 1, this was accomplished by doing three center-point runs, or, in other words, three runs at target conditions for each factor. Note that this also provided an estimate for curvature for the response data (see the ANOVA results in Fig. 2). However, a better way to get an error estimate for the design matrix is to replicate the experiment. As it is very easy to set up the station, perform the experimental runs, and test the parts, we will get a good error estimate for this study by replicating the DOE five times. That means each experimental run listed in Table 3 will be run five times. Thus, the complete DOE matrix with all 45 runs is given in Table 4. This does not mean that

**Table 4** Design Matrix for Case Study 2 with Five Replications and Response Data

| Standard order | Run order | Time (sec) | Pressure (psi) | Vacuum level | Bond strength (lbs) |
|---|---|---|---|---|---|
| 1 | 12 | 0.2 | 10 | 5 | 24.10 |
| 2 | 1 | 0.4 | 10 | 5 | 26.15 |
| 3 | 5 | 0.2 | 30 | 5 | 26.80 |
| 4 | 20 | 0.4 | 30 | 5 | 25.85 |
| 5 | 41 | 0.2 | 10 | 15 | 25.55 |
| 6 | 16 | 0.4 | 10 | 15 | 27.05 |
| 7 | 8 | 0.2 | 30 | 15 | 24.70 |
| 8 | 2 | 0.4 | 30 | 15 | 25.50 |
| 9 | 38 | 0.2 | 10 | 5 | 18.50 |
| 10 | 15 | 0.4 | 10 | 5 | 28.65 |
| 11 | 45 | 0.2 | 30 | 5 | 24.95 |
| 12 | 31 | 0.4 | 30 | 5 | 25.95 |
| 13 | 26 | 0.2 | 10 | 15 | 23.05 |
| 14 | 40 | 0.4 | 10 | 15 | 26.55 |
| 15 | 21 | 0.2 | 30 | 15 | 24.70 |
| 16 | 28 | 0.4 | 30 | 15 | 24.45 |
| 17 | 23 | 0.2 | 10 | 5 | 28.70 |
| 18 | 3 | 0.4 | 10 | 5 | 24.00 |
| 19 | 9 | 0.2 | 30 | 5 | 25.30 |
| 20 | 17 | 0.4 | 30 | 5 | 25.05 |
| 21 | 7 | 0.2 | 10 | 15 | 26.30 |
| 22 | 36 | 0.4 | 10 | 15 | 24.30 |
| 23 | 30 | 0.2 | 30 | 15 | 26.80 |
| 24 | 33 | 0.4 | 30 | 15 | 26.20 |
| 25 | 14 | 0.2 | 10 | 5 | 25.30 |
| 26 | 25 | 0.4 | 10 | 5 | 24.45 |

(*Continued*)

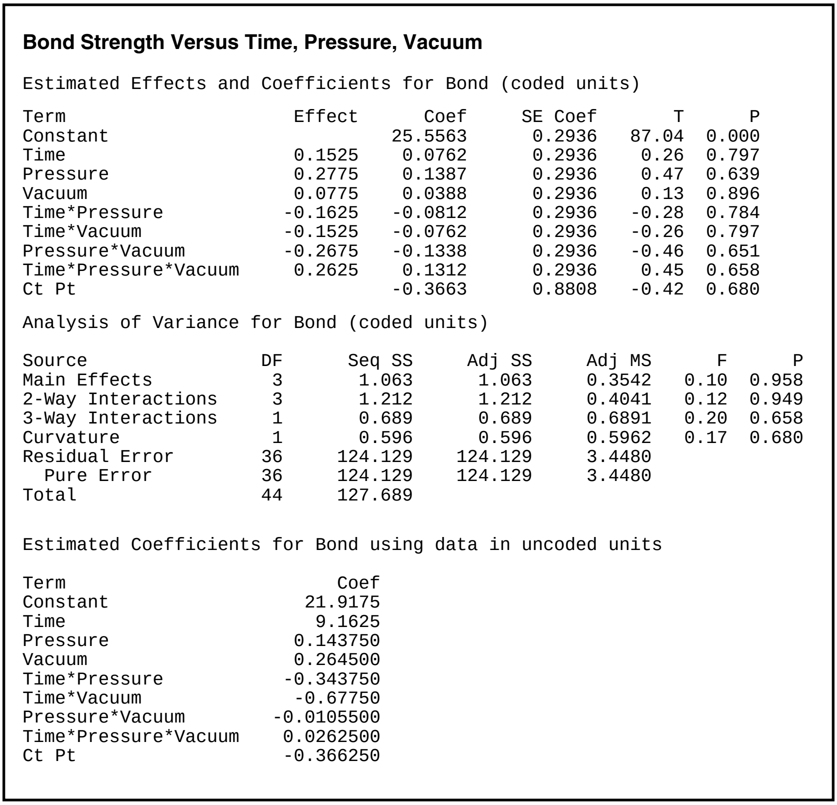
**Table 4** Design Matrix for Case Study 2 with Five Replications and Response Data (*Continued*)

| Standard order | Run order | Time (sec) | Pressure (psi) | Vacuum level | Bond strength (lbs) |
|---|---|---|---|---|---|
| 27 | 22 | 0.2 | 30 | 5 | 27.70 |
| 28 | 6 | 0.4 | 30 | 5 | 25.75 |
| 29 | 39 | 0.2 | 10 | 15 | 24.45 |
| 30 | 42 | 0.4 | 10 | 15 | 24.10 |
| 31 | 18 | 0.2 | 30 | 15 | 26.45 |
| 32 | 27 | 0.4 | 30 | 15 | 26.45 |
| 33 | 29 | 0.2 | 10 | 5 | 27.80 |
| 34 | 24 | 0.4 | 10 | 5 | 24.80 |
| 35 | 10 | 0.2 | 30 | 5 | 24.50 |
| 36 | 11 | 0.4 | 30 | 5 | 26.05 |
| 37 | 37 | 0.2 | 10 | 15 | 28.85 |
| 38 | 4 | 0.4 | 10 | 15 | 25.70 |
| 39 | 35 | 0.2 | 30 | 15 | 25.10 |
| 40 | 19 | 0.4 | 30 | 15 | 25.65 |
| 41 | 43 | 0.3 | 20 | 10 | 25.80 |
| 42 | 34 | 0.3 | 20 | 10 | 26.65 |
| 43 | 44 | 0.3 | 20 | 10 | 24.30 |
| 44 | 32 | 0.3 | 20 | 10 | 23.80 |
| 45 | 13 | 0.3 | 20 | 10 | 25.40 |

all DOE studies require five replications. But, it is a very good idea (when possible) to replicate and confirm your experiments to make sure your conclusions are accurate. This point will be clearly illustrated later in this chapter.

*Step 5*: Determine the sample size for each run.

The sample size for each individual run will be one part. However, as we are performing each run five times, we will actually end up with five variable data points for each of the nine different combinations of factors. According to MINITAB, by replicating the DOE five times, we will have a very high likelihood of detecting the magnitude of "effect" that we are looking for. The number of replications is determined using the "Power and Sample Size" function for a two-level factorial design in MINITAB.



**Bond Strength Versus Time, Pressure, Vacuum**

```
Estimated Effects and Coefficients for Bond (coded units)

Term                    Effect      Coef     SE Coef       T       P
Constant                         25.5563      0.2936   87.04   0.000
Time                    0.1525    0.0762      0.2936    0.26   0.797
Pressure                0.2775    0.1387      0.2936    0.47   0.639
Vacuum                  0.0775    0.0388      0.2936    0.13   0.896
Time*Pressure          -0.1625   -0.0812      0.2936   -0.28   0.784
Time*Vacuum            -0.1525   -0.0762      0.2936   -0.26   0.797
Pressure*Vacuum        -0.2675   -0.1338      0.2936   -0.46   0.651
Time*Pressure*Vacuum    0.2625    0.1312      0.2936    0.45   0.658
Ct Pt                            -0.3663      0.8808   -0.42   0.680

Analysis of Variance for Bond (coded units)

Source                DF      Seq SS      Adj SS      Adj MS       F       P
Main Effects           3       1.063       1.063      0.3542    0.10   0.958
2-Way Interactions     3       1.212       1.212      0.4041    0.12   0.949
3-Way Interactions     1       0.689       0.689      0.6891    0.20   0.658
Curvature              1       0.596       0.596      0.5962    0.17   0.680
Residual Error        36     124.129     124.129      3.4480
  Pure Error          36     124.129     124.129      3.4480
Total                 44     127.689


Estimated Coefficients for Bond using data in uncoded units

Term                       Coef
Constant                21.9175
Time                     9.1625
Pressure               0.143750
Vacuum                 0.264500
Time*Pressure         -0.343750
Time*Vacuum            -0.67750
Pressure*Vacuum      -0.0105500
Time*Pressure*Vacuum  0.0262500
Ct Pt                 -0.366250
```

**Figure 8** MINITAB™ print-out for case study 2.

*Step 6*: Perform the experiments/compile the data.

After running each of the 45 experiments given in Table 4, the data is added to the design matrix and analyzed using MINITAB. The actual response variable data for bond strength is also given in Table 4.

*Step 7*: Analyze the data/draw conclusions.

The MINITAB analysis Printout is given in Figure 8. The pareto chart (Fig. 9) and corresponding normal probability plot (Fig. 10) show that there are no statistically significant main effects or interactions present. Note that the bars are all to the left of the verticle line on the pareto chart and none of the points on the probability plot are labeled. In addition, looking at the data

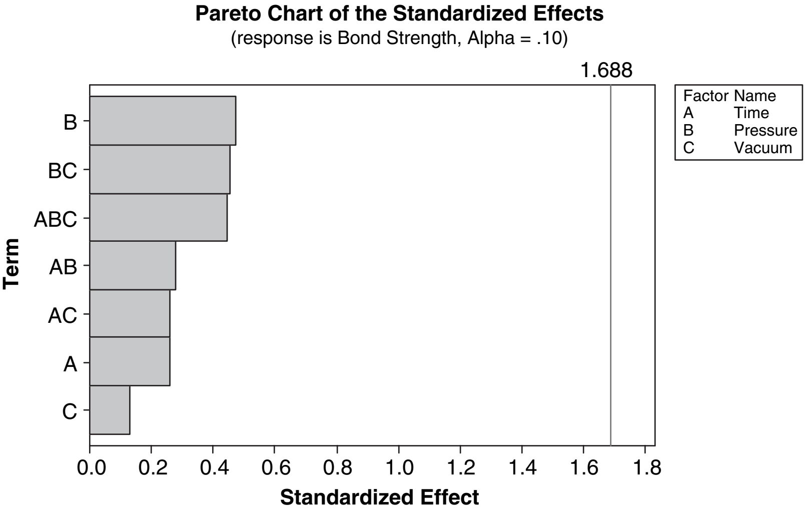


**Figure 9** Pareto chart for case study 2.

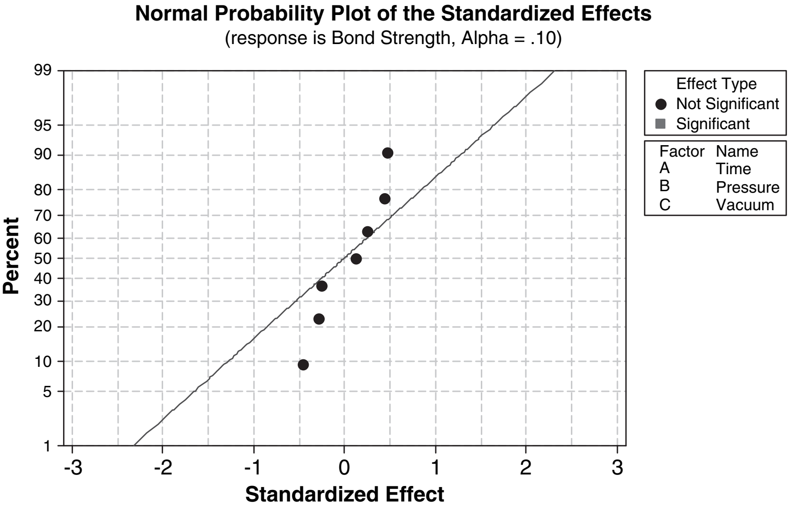


**Figure 10** Normal probability plot for case study 2.

from a position of practical significance, the calculated effects in the ANOVA table (Fig. 8) are extremely low numbers. The largest individual effect (the pressure) is less than 0.3 lbs. Considering the average bond strength for the 45 runs is 25.5 lbs and the lowest reading was 18.5 lbs, a fluctuation of 0.3 lbs seems hardly significant (especially when the product specification for bond strength is only 10 lbs). Some may look at this data and think that we have not learned anything because nothing affects the response variable. However, quite the opposite is true. We have learned that, over the ranges studied, none of the three controllable factors (time, pressure, vacuum level) significantly affect the bond strength of the finished product. This is a robust process! This is exactly the kind of data that we want to see prior to validation studies and future commercial manufacturing.

*Step 8*: Eliminate or reduce aliases.

As the DOE is a full-factorial design, there are no aliases to worry about.

*Step 9*: Confirm the results.

As we have already performed five replications of a full-factorial experiment, we can feel pretty good about our results. In addition, we now know that the three factors of time, pressure, and vacuum level do not affect the response variable (bond strength) over the ranges studied. Therefore, it really does not make sense to repeat the full-factorial as a part of a validation study. In this case, it would probably make sense to proceed directly to process validation by doing three runs (per approved protocol) at target values for time, pressure, and vacuum level. Ideally, this work could just be included as part of the final product qualification, where the manufacturing line is run for three separate eight-hour shifts, and the finished product from each shift is sampled and tested per the protocol. Of course, the rationale for the scope of the validation work must be clearly documented (usually in a "Background" section in the protocol) and supported by the data that is compiled and discussed in a well-written Development Report.

Finally, let's conclude this chapter with an extremely important discussion of statistics and the value of data replication and confirmation. We all know that all statistics have an associated probability that goes along with them. Without going into a long discussion on the subject, what this means to us is that there is always a chance that a wrong conclusion may be drawn from a given data set. There is always a chance that we may obtain some bad data or even a statistical outlier in our final response data. This is especially true for small sets of data.

When we use DOE, the possibility of drawing a wrong conclusion is greatest when the experimental matrix is not replicated at least once. Naturally, the more data you have, the more likely you are to gain an understanding of the "truth." When doing a designed experiment, once again we need to put our brains into action and think about what our results are telling us. In case study 1, the matrix was not replicated. In other words, we did not repeat every experimental run. However, the results did not deviate from what was expected, and they certainly made sense from a logical standpoint (i.e., no surprises). This is typical for a DOE performed early in the process development stage. There may be several factors included in the design matrix that are, based on historical data or scientific rationale, probably not significant, but must be checked out anyway just to be sure. Thus, it is often possible to get by without replicating a DOE that is performed early in process development. And let's face it, if you have eight or more factors in your design matrix, the amount of time and effort required to complete all of the experiments just once may already be taxing your resources (even without doing a replication). It is also important to note, for this particular example, that we did take 568 samples per run in order to calculate a fraction defective. So we did have more than one sample per run, but only one variable data point.

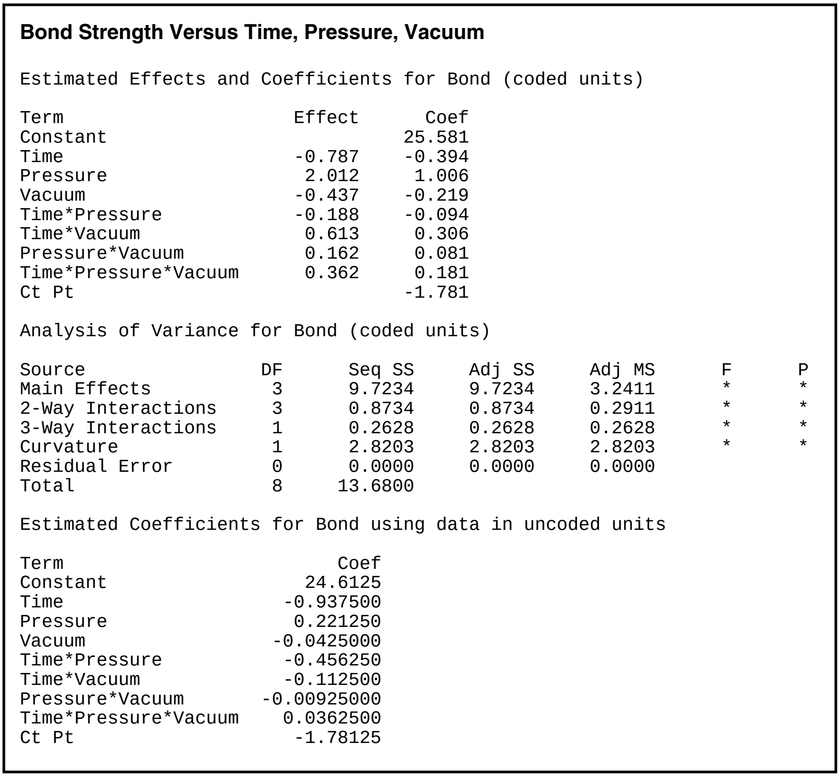
Case study 2 was performed just prior to the final process/product validation studies. There were only three controllable factors to consider that might have an effect on the response variable. However, there was a good degree of uncertainty regarding which, if any, were significant. In this case study, the previous process development work was limited at best. In addition, the required experiments were easy and inexpensive to perform. Therefore, it made a lot of sense to plan for replications up front. Remember, we always want to be highly confident that we understand the process prior to going into validation studies. It probably was not necessary to perform five replications of the design matrix, but usually three is a good number to pursue when possible.

In order to provide a clear example of the danger of not replicating a DOE, let's take a look back at the data for case study 2. By performing five replications, it was ultimately clear that, over the ranges studied, none of the three factors had a significant impact on the response variable (bond strength). But what if we had not replicated the DOE? Is it possible that we could have got different results? Of course, the answer to that question is a resounding yes!

Hence, let's see what happens when we take just one of the five replicated data sets and analyze it using MINITAB. The data will be from the fourth replication of the DOE and is given in Table 5. Using our same method of data analysis as earlier, we interpret the data using the graphs and ANOVA print-out given in Figures 11–13. Clearly, this data is saying

**Table 5** Design Matrix and Response Data for Case Study 2 (Fourth Replication Only)

| Standard order | Time (sec) | Pressure (psi) | Vacuum level | Bond strength (lbs) |
|---|---|---|---|---|
| 1 | 0.2 | 10 | 5 | 25.30 |
| 2 | 0.4 | 10 | 5 | 24.45 |
| 3 | 0.2 | 30 | 5 | 27.70 |
| 4 | 0.4 | 30 | 5 | 25.75 |
| 5 | 0.2 | 10 | 15 | 24.45 |
| 6 | 0.4 | 10 | 15 | 24.10 |
| 7 | 0.2 | 30 | 15 | 26.45 |
| 8 | 0.4 | 30 | 15 | 26.45 |
| 9 | 0.3 | 20 | 10 | 23.80 |



**Bond Strength Versus Time, Pressure, Vacuum**

```
Estimated Effects and Coefficients for Bond (coded units)

Term                     Effect      Coef
Constant                           25.581
Time                     -0.787    -0.394
Pressure                  2.012     1.006
Vacuum                   -0.437    -0.219
Time*Pressure            -0.188    -0.094
Time*Vacuum               0.613     0.306
Pressure*Vacuum           0.162     0.081
Time*Pressure*Vacuum      0.362     0.181
Ct Pt                              -1.781

Analysis of Variance for Bond (coded units)

Source                 DF      Seq SS      Adj SS      Adj MS       F       P
Main Effects            3      9.7234      9.7234      3.2411       *       *
2-Way Interactions      3      0.8734      0.8734      0.2911       *       *
3-Way Interactions      1      0.2628      0.2628      0.2628       *       *
Curvature               1      2.8203      2.8203      2.8203       *       *
Residual Error          0      0.0000      0.0000      0.0000
Total                   8     13.6800

Estimated Coefficients for Bond using data in uncoded units

Term                         Coef
Constant                  24.6125
Time                    -0.937500
Pressure                 0.221250
Vacuum                 -0.0425000
Time*Pressure           -0.456250
Time*Vacuum             -0.112500
Pressure*Vacuum       -0.00925000
Time*Pressure*Vacuum    0.0362500
Ct Pt                    -1.78125
```

**Figure 11** MINITAB™ print-out for case study 2 (fourth replication only).

that factor B (pressure) is statistically significant in affecting the bond strength of the finished part. One could still argue that the practical significance is low, but the statistical conclusions to be drawn from this limited data set are clearly not accurate.

In summary, DOE is an exceptionally powerful and efficient method for planning, running, and analyzing experiments during process/product development. It is used to determine which process factors are insignificant so that they can be eliminated from future studies. In addition, it provides a quantitative estimate of the effect that a factor has on one or more response variables. It also provides the same quantitative estimate for interactions between factors.

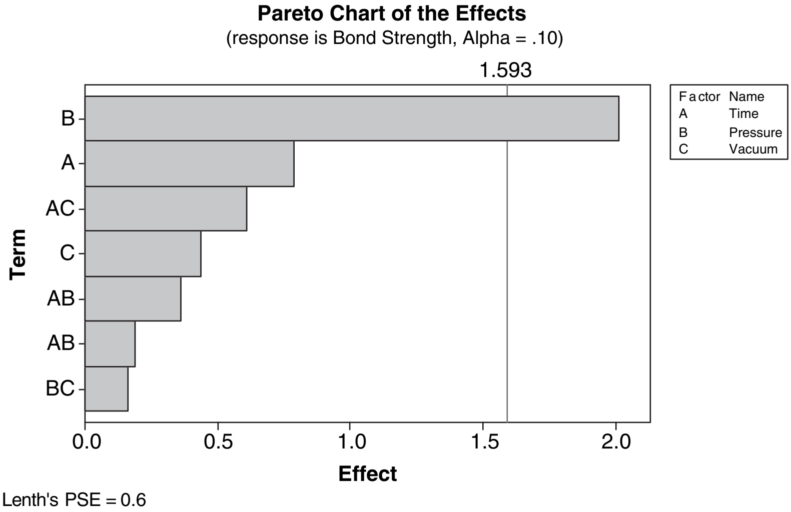


**Figure 12** Pareto chart for case study 2 (fourth replication only).

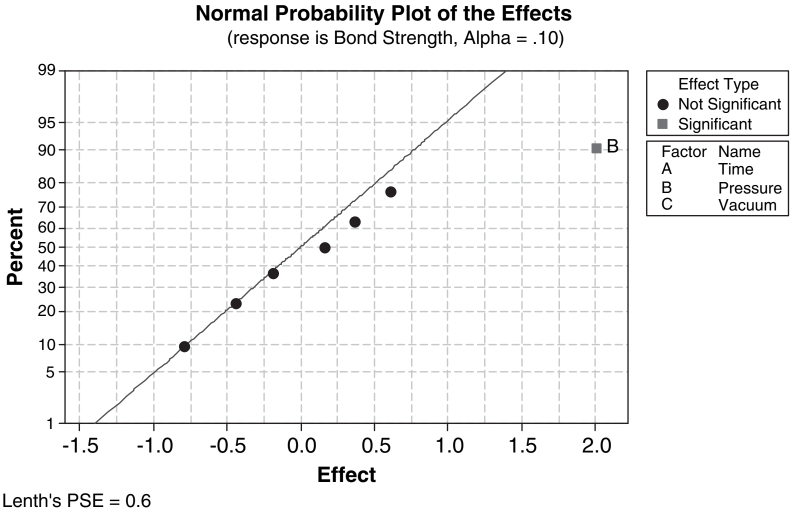


**Figure 13** Normal probability plot for case study 2 (fourth replication only).

This is often times very important for pharmaceutical production processes where interactions among factors are common. Ultimately, DOE allows the experimenter to determine which process factors are to be included in subsequent validation studies. By eliminating the insignificant factors from the validation work, a great deal of time and money can often be saved. In addition, by using DOE to streamline the development work and efficiently characterize the process, the timeline for product development and a successful commercial launch can be shortened significantly.

## ACKNOWLEDGMENT

Special thanks to Dennis Bialecki for his contributions to the second case study.

# Index

Pharmaceutical Science

## about the book...

This title demonstrates how designed experiments are the most scientific, efficient, and cost effective method of data collection for validation. Intended as a learn-by example guide, ***Pharmaceutical and Medical Device Validation by Experimental Design*** demonstrates why designed experiments are the most logical and rational approach to use, using realistic case studies, illustrations, and where appropriate, step-by-step protocols and procedures.

This title shows the function of designed experiments for process analytical technologies (PAT)...illustrates the wide-range of applications for designed experiments...demonstrates the importance of well-organized process development work to ensure a successful validation study.

## about the editor...

LYNN D. TORBECK is an International Consultant specializing in applied statistics and experimental design for pharmaceutical and biopharmaceutical development, quality assurance, and control validation, Torbeck & Associates, Inc., Evanston, Illinois. Mr. Torbeck received the B.S. and M.S. degrees in Statistics, University of Tennessee, Knoxville, and formerly was Director of Validation World-Wide Quality Assurance, G. D. Searle (Pfizer), Skokie, Illinois.

*Printed in the United States of America*

H5569

ISBN 978-142005569-6

90000

9 781420 055696

informa
healthcare
www.informahealthcare.com

52 Vanderbilt Avenue
New York, NY 10017

Telephone House
69-77 Paul Street
London EC2A 4LQ, UK