CHAPTER

Application of quality by design (QbD) to the development and validation of analytical methods

3

David K. Lloyd*, James Bergum†

* *Analytical and Bioanalytical Development, Bristol-Myers Squibb, New Brunswick, NJ, USA,*
† *Statistical Consultant, BergumSTATS, LLC, Howell, NJ, USA*

CHAPTER OUTLINE

Specification of Drug Substances and Products. http://dx.doi.org/10.1016/B978-0-08-098350-9.00003-5

Copyright © 2014 Elsevier Ltd. All rights reserved.

3.1 INTRODUCTION

3.1.1 Analytical quality by design

Quality by design (QbD) has been proposed as a systematic approach to product development, wherein understanding of the product is paramount.[1,2] Through a comprehensive understanding of the effects of various inputs (e.g. process parameters, materials) on the final product (active pharmaceutical ingredient or drug product), appropriate ranges of the input parameters may be defined within which the quality of the final product is guaranteed.[1,2] Achieving the appropriate level of understanding typically involves a multifactor approach to the study of a product, since changes to one input may alter the effect of another input on the final product. One-factor-at-a-time (OFAT) studies are simply not adequate to develop broad product understanding. On the other hand, given the large number of potential input parameters for any product, development of a multifactor model for all possible effects is practically impossible. Therefore, risk analysis is an important element of QbD,[3] allowing the truly impactful parameters to be identified for further investigation in studies which are of a manageable scale. These studies generally apply statistical design of experiments (DOE) or mechanistic models, and the output is a multidimensional design space in which the effects of key parameters are understood, and a product control space is defined where appropriate product quality is guaranteed (or at least highly probable!) within the parameter ranges that border the space. A further important element of QbD is the development of a control strategy, which governs how changes are assessed and implemented during the life of a product.[4]

Analogous to QbD for product, QbD may also be applied in the analytical realm.[5] If QbD for a product is defined as a full understanding of how inputs and process attributes relate to product performance, then for analytical methods it can be considered to be a full understanding of how the analytical technique attributes and operating conditions relate to analytical performance. Factors that may be considered for study include the type of analytical technique chosen, reagents used, and instrumental parameters. The method performance is defined by both the type of data that the method produces and the required quality of that data. Data quality has traditionally been determined at the *grande finale* of method development process, method validation, when method performance is determined (accuracy, precision, selectivity, robustness, etc.). By applying QbD to analytical methods, the method performance is instead largely defined and understood during the development process.

3.1.2 Why do it?

In the authors' experience, raising the topic of QbD to an analytical audience results in, at best, a mixed reaction (indeed, this may be the case for nonanalytical audiences as well[6]). Typically, concerns are raised as to what the analyst will get out of it (what benefits will be gained), and even more so about how much extra work QbD will require. One potential benefit that is sometimes proposed is that by defining a QbD design space, the method could be operated anywhere within that design space. Although this has merit in providing some flexibility in manufacturing processes, it is less obvious that this represents a major advantage in a typical analytical method. Although from the design space you may know that you can perform your measurement at a wavelength of ±10 nm from the method set point of 254 nm, it's unlikely that there would be much value in deviating from the set point; "It's Tuesday, I think I'll try 264 nm today…" is probably not a behavior typical of most analytical scientists. Instead, the great benefit comes from the identification of a robust operating region for the method. Method failure due to lack of robustness can have considerable impact, e.g. delaying a project start due to a failed technology transfer, or imperiling batch release if there is a method failure during product testing. If analytical QbD can help assure clean method transfers and routine smooth method operation, the long-term impact and value will be high. As to the question of how much extra work analytical QbD entails, the answer probably ranges from "little or nothing" to "a lot", depending on how well QbD is built into the method development and validation process. If it comes as an afterthought, it will surely result in extensive extra work. If QbD is built into the process from the beginning, good risk assessment is performed to eliminate low-value studies, and the results of systematic method development are contemporaneously documented, the impact on time and effort should be minimal while increasing method understanding and robustness.

3.2 METHOD REQUIREMENTS

The requirements for a particular analytical method are strongly tied to the product and the product attribute to which the method is being applied. The actual measured result will contain elements of method and product variability combined via their variances (see also Section 6.2.4.2 in Chapter 6):

$$\sigma^2 = \sigma^2{}_{\text{product}} + \sigma^2{}_{\text{method}} \tag{3.1}$$

It is desirable to reduce the method variability such that it becomes a relatively small contribution to the overall variability. In addition, the method may suffer from other errors such as bias, interference, etc., which reduce the data quality. The method requirements should be set such that the data generated by the method are a good reflection of the product attribute being tested, and not masked by analytical errors.

With some types of analyses, it is relatively easy to ensure that the errors of analysis are small relative to the variability of the product being measured. For example, in a chromatographic test for content uniformity (CU), it would typically be unusual for the method variance to be more than a small fraction of the product variance. With relatively cursory method development, the method requirements are likely to be met without a full-blown application of QbD. However, for other measurements the potential for analytical error may be relatively large and there is correspondingly greater justification for more extensive systematic studies.

Required method characteristics may be defined in an analytical target profile (ATP), which may be viewed as being somewhat similar to a specification for an analysis,[7] and analogous to a quality target product profile.[2] The ATP lists important method characteristics (e.g. parameters such as accuracy, precision) and describes the degree to which these must be controlled (e.g. what percentage of inaccuracy or imprecision is acceptable). Note that the ATP is essentially independent of the analytical technique used; it simply defines the characteristics that the method must have in order to adequately measure the product's critical quality attributes (CQAs). The technique-independent nature of the ATP was originally envisaged as offering a way to include required method characteristics in a regulatory filing without actually specifying exact method conditions; any method could be used to measure a product CQA so long as it was demonstrated to meet the ATP. In our experience this is not yet a concept which is broadly embraced by regulatory authorities; however, it remains a useful tool in defining what a method has to measure and how well it has to make that measurement. Given a defined ATP, one can then decide on how to fulfill the ATP's requirements, in other words, what sort of method to use. When multiple techniques offer the required analytical performance, factors not related to data quality such as cost, speed and "greenness" become important. In some cases the choice is fairly obvious, e.g. a large majority of small molecule impurity analyses are performed by reversed-phase liquid chromatography because it is a relatively routine, inexpensive technique, which is well suited to meet the ATP's requirements for typical small drug molecules.

3.3 METHOD RISK ASSESSMENT

3.3.1 Definition of risk

From ICH Q9,[3] "risk" is defined as "The combination of the probability of occurrence of harm and the severity of the harm." Risk can mean many different things in the context of pharmaceutical development, but from a viewpoint of regulatory agencies this is principally the risk of harm to the patient. In the context of analytical testing one can consider the possible harm being caused to the patient through an incorrect analytical result leading to release of a batch with undesirable characteristics. In the context of pharmaceutical analysis it has been proposed that, using a failure mode effects analysis approach (Section 3.3.2) risk = severity × occurrence × detectability, where these terms are defined by Nasr[8] as follows:

- severity = effect on patient
  - related to safety or efficacy
- likelihood of occurrence = chance of failure
  - related to the quality and extent of product and process knowledge and controls
- detectability = ability to detect a failure
  - related to suitability of the analytical methodology (sampling and testing)

It is important to note that the above expression is not a robust quantitative mathematical relationship, but a more qualitative statement that something is high risk if it has bad consequences, is likely to happen, and is not likely to be detected. The definition of "severity" may usefully be expanded; from a business perspective, harm could come about because of incorrect analytical results leading to rejection of a truly good batch of product. Thus we may choose to include business risks in the calculation (although the regulatory perspective is likely to be narrower, excluding these from consideration).

The process of risk assessment for analytical methods is thus a determination (qualitative or quantitative) of the effects of variation in factors such as method operating parameters or sample characteristics on method performance. Assessment of risk includes identification of potential risks, and analysis of these risks leading to an evaluation of the importance of that risk. The risk-assessment process brings an important benefit to method development, because a good analysis of what are truly important parameters allows a more focused systematic study of only those parameters.

3.3.2 Risk assessment toolbox

An element of risk assessment is present in any thoughtful method development activity. Traditionally, it may not have been performed as a separate activity, but informally (maybe just in the analyst's head!) as a consequence of the analyst's general knowledge of the technique and sample at hand. However, a more formal approach allows the risks to be documented and decisions justified. There exist many tools which aid in structuring the risk assessment process:

- Qualitative tools for parameter screening, e.g.
  - Process mapping
  - Ishikawa or fishbone diagrams

Such qualitative tools help define risks in a process or method by systematically laying out the various method steps and identifying the associated risks. For example, the process of a drug product analysis may be mapped as involving an automated sample preparation followed by a chromatographic analysis. These two steps in the process can each be further broken down into sub-operations, sub-sub-operations, and so on. Possible elements of risk can be associated with each operation. Fishbone diagrams illustrate the process somewhat differently (see Section 3.3.3 for an example). Many risk factors could be identified with each process, but combining one's general analytical knowledge about the technique in use and specific knowledge about the particular analyte, many hypothetical risks can quickly be discounted, leaving relatively few potentially critical parameters for further consideration.

- Semiquantitative tools for risk ranking, e.g.
  - Relative ranking
  - Failure mode effects analysis/Failure mode effects and criticality analysis (FMEA/FMECA)

Semiquantitative tools for risk ranking help define which elements of the method are truly CQAs. For example, after initial qualitative triage of risk factors, the remaining parameters can be assessed for their relative criticality; a factor that has potential to result in erroneous data which cannot easily be identified as erroneous would rate as high risk, whilst one where the error can easily be spotted would rate lower. FMEA/FMECA is specifically mentioned in ICH Q9 "Quality Risk Management".[3] FMEA starts by evaluating potential failure modes for each step in the analytical method, and correlates each failure mode with a likely effect either for patient safety or as a business risk. The root cause for each failure mode is postulated based on previous knowledge or general scientific principles. The assessment may be extended to include a consideration of the criticality of a particular risk, and hence may allow identification of method steps where additional preventive actions may be appropriate to minimize risks. It should be emphasized that although a risk probability number (RPN) may be assigned as the product of severity, likelihood and detectability (see Section 3.1 above), the RPN is not a hard number in the sense that, for example, one expects an assay value to be. Although there are criticisms of the FMEA methodology,[9,10] it nevertheless provides a useful framework within which to attempt to systematically identify and rank risks and as such, can be useful provided one does not attempt to over-interpret the numerical output.

- Experimental tools for process understanding, e.g.
  - Statistically designed experiments
  - Mechanistic models

Statistical tools can support and facilitate risk assessment. For example, a screening factorial design can demonstrate the sensitivity of a method to different parameters, and the variability encountered within the experimental space. Mechanistic models may provide similar information—a simple univariate example would be the Henderson–Hasselbach equation relating the degree of ionization of a compound ([salt]/[acid]) to its pK_a and the solution pH:

$$\mathrm{pH} = \mathrm{p}K_\mathrm{a} + \log([\mathrm{salt}]/[\mathrm{acid}]) \tag{3.2}$$

The effects of the relationship described by this equation on chromatographic method robustness are enshrined in the chromatographer's axiom that one should operate >2 pH units away from the analyte's pK_a so that minor changes in pH will not cause a large change in the ratio [salt]/[acid] (these species typically having very different retention). In the common situation of chromatographic separation of impurities where species with varying pK_as are simultaneously analyzed, an understanding of the charge state of each component will illustrate which, if any, are potentially at risk of varying retention from pH change and thus whether pH should be considered a primary factor for further investigation.

It can be very helpful to perform risk identification as a small-group activity. Having multiple viewpoints in the risk assessment reduces the possibility that potential risks will be overlooked or dismissed (the analyst considering Eqn (3.2) may believe that the analyte pK_as are far from the operating pH, but hopefully in a group discussion someone would remember that the pK_as of weak acids and bases depend greatly on the solvent mixture used and ask whether the "known" values are correct under the chromatographic conditions used!). The assessment should be appropriately comprehensive, and not just limited to the method conditions. For example, the sample itself is an important risk factor in many analyses; if there are variations in tablet properties such as particle size or hardness, these may well affect a spectroscopic calibration model or an extraction. It may well be that potential risks are identified that cannot be fully addressed at a given point in development; although analyst-to-analyst variability may be studied in a single lab, an investigation of lab-to-lab variability may not be possible (or warranted) in early development (see also Chapters 4 and 6). In later development, a wide variety of samples from product development QbD studies may become available, which cover the extremes of the process operating ranges. These may not be ready for inclusion in method development studies, but the potential risk can be identified early, and the effect (or, hopefully, lack of effect) confirmed later, e.g. in a separate ruggedness study.

3.3.3 Risk assessment example

Rather than starting with an example from a specific analytical technique, consider an example which almost any analytical chemist can relate to: the commute to work. Figure 3.1 shows a fishbone diagram

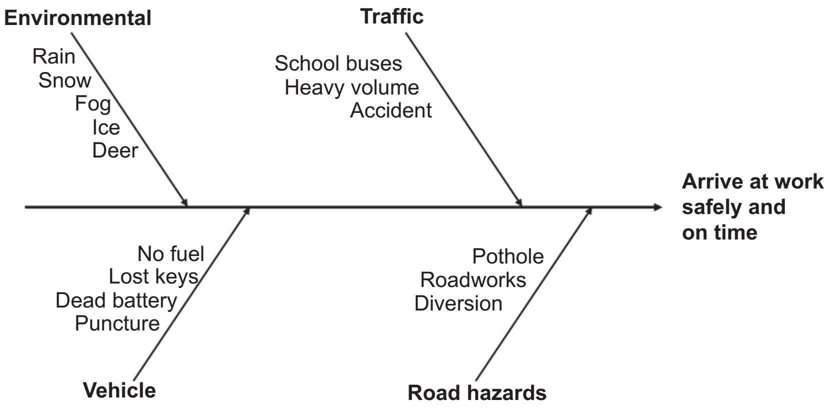

FIGURE 3.1

Fishbone diagram illustrating various risks involved in the daily commute to work. The goals are identified to the right: arriving at work safely and on time.

where four major categories of risk have been identified, with several specific risks shown in each category. Note that to the right, the goals of the commute are defined—to arrive at work safely and on time.

A qualitative assessment of the risks indicates that several can be avoided by appropriate planning. For example, a good standard operating procedure (SOP) on vehicle maintenance would likely eliminate the risks of being delayed due to a dead battery or a puncture, while another related to vehicle operation should ensure that the commuter appropriately fuels their vehicle and systematically stores their keys in a place they will not be lost. Thus, systems can be put in place to greatly minimize the identified vehicle risks. The traffic risks can also be mitigated to a significant extent, e.g. an early commute avoids volume and school buses, while forward planning can avoid major roadworks. In Table 3.1, a more quantitative analysis of these risks is presented.

The risks are rated in terms of their possible effect on achieving the stated goals. These goals need not be equally weighted when considering the risks; hopefully safety will have a greater weight in the risk analysis than timeliness. Thus, although losing one's keys in the morning may delay arrival, this is a relatively minor inconvenience compared to an unplanned close encounter with a deer (rated high severity) and was given a lesser weight for severity in the risk assessment. Similarly, judgments are made on the likelihood of a hazard occurring. If vehicle-related SOPs are in place, vehicle-related problems should have a low probability. Finally, there is the question of detectability. A hazard which can easily be detected can be avoided, and hence is given a low rating. On the other hand, black ice or a deer hiding in the roadside woods carries a higher detectability rating (i.e. poor detectability). The overall risk rating is the product of the severity, likelihood and detectability ratings. From the above analysis, it is estimated that the environmental hazards of ice and deer are the critical hazards which still need to be addressed for this commute as a result of their severity and poor detectability. These are identified as primary factors, while five other factors have a more modest impact and may optionally be assessed further.

Table 3.1 FMEA analysis of a daily commute

Risk	Severity High = 3 Medium = 2 Low = 1	Likelihood High = 3 Medium = 2 Low = 1	Detectability High = 1 Medium = 2 Low = 3	Numerical Rating Detectability × Likelihood × Severity	Primary Factor?
Environmental					
Rain	1	2	1	2	N
Snow	2	1	1	2	N
Fog	2	1	1	2	N
Ice	3	1	3	9	Y
Deer	3	1	3	9	Y
Vehicle					
No fuel	1	1	1	1	N
Lost keys	1	1	3	3	N
Dead battery	1	1	2	2	N
Puncture	2	1	2	4	?
Traffic					
School buses	1	2	1	2	N
Heavy volume	1	2	2	4	?
Accident	2	1	2	4	?
Road hazards					
Pothole	1	2	2	4	?
Roadworks	1	2	2	4	?
Diversion	1	2	1	2	N

The quality of the risk analysis will impact the factors studied going forward, and thus the extent of work that will be done. In this case, the hazards were identified in a brainstorming session, and the ratings were made subjectively. One can see that this process could be improved upon; for example, real statistics for road accidents or traffic flow could have been sought to better quantify the hazards and potentially to identify ones which were not recognized.

3.4 METHOD DEVELOPMENT AND OPTIMIZATION: UNDERSTANDING THE METHOD OPERATING SPACE

With many analytical techniques, one can achieve useful results with practically no method development or optimization. If rather standard approaches and method parameters give an acceptable result for the large majority of samples, and sources of error with the technique are well understood, extensive method development and optimization studies are probably not a good use of resources. Tests such as Karl Fischer (KF), simple UV measurements, or some compendial methods may fall into this category. On the other hand, there are many analyses where extensive development and optimization experiments are the norm, e.g. chromatographic impurity analyses, dissolution, or particle-size measurements. Even with these techniques, one can be fortunate and achieve a reasonable result after a few experiments based on very generic conditions; a broad acetonitrile–water gradient on a C18 column is a standard starting point for small molecule pharmaceutical analysis and it is not so unusual for many components in a sample to be resolved on the first attempt. However, this is not the whole story since the ATP will contain a variety of quality attributes such as accuracy and precision, and the choice should also include business requirements related to analytical speed or greenness of the method. Further experimentation is required to determine appropriate conditions where the ATP is met with good method robustness. The goal of these experiments is to understand the effect of the previously identified primary factors affecting the method.

There are a variety of ways in which experimental data can be transformed into useful method knowledge. Trial-and-error and more systematic OFAT approaches are limited because they provide information about points or lines in experimental space, but cannot be interpreted to understand method behavior across large regions of the experimental space. However, empirical models of the method space can be built using appropriately designed multifactor experiments, which are amenable to interpretation in a way that OFAT experiments are not. DOE approaches such as factorial designs or response surface designs fit responses to empirical functions, e.g. a quadratic function including cross terms. Such models are not intended to be expressions of the physico-chemical processes underlying the analytical method, but they do allow a result to be predicted based on a combination of input factors. Because they are not built around a method-specific model, they are universally applicable, albeit at the cost of requiring a significant number of experiments to build the model. A variety of DOE approaches useful for analytical QbD are described in Section 3.5.

In some cases, an explicit mechanistic model is available, which describes the analytical process based on a fundamental understanding of the technique. For example, in chromatography, a number of commercial software packages are available that are built around theoretical descriptions of the chromatographic separation. The advantage of this sort of approach is that the analytical response can in many cases be accurately modeled based on a very few, carefully chosen experiments. Such approaches are discussed in Section 3.6.

3.5 EMPIRICAL MODELS: DOE (SCREENING, MODELING, ROBUSTNESS)

3.5.1 Introduction

Use of an empirical model founded on statistically based DOE is a powerful tool in QbD method development. There are various software packages such as JMP, SAS, Design-Expert and Minitab that can generate a design and/or analyze the results. A DOE not only provides an organized approach to problem solving but also enables efficient and clear estimation of the effects that factors have on responses. Factors (independent variables) are chosen and controlled by the experimenter. Factors can be qualitative such as column type or solvent that are generally called “class” factors and are not on a continuous scale, or quantitative such as time or speed that are generally called “continuous” factors. Each factor is studied at one or more levels. For example, the factor may be solvent but there may be four different solvents studied. The four solvents are the levels of the solvent factor. The factor could be quantitative such as speed or time with levels of 4, 6, and 8 rpm or 1, 3, and 5 min, respectively. Responses (dependent variables) such as %recovery, %residual solvent, potency (mg/tablet) are the measured results that are generated from application of the factors to experimental material. Once the factors and levels are chosen, there are two primary components in constructing the design: (1) the combination of factor levels to include in the design and (2) the number of times to replicate each combination of factor levels. Of course, there are many details to consider prior to addressing these questions, such as: What is the question which will be answered by conducting the study? What are the available resources (materials, machines, and people)? What are the constraints on the factor levels? How will the design be carried out? What is the current available knowledge? What are the known issues about the factors and/or responses? These and other questions are addressed by definition of the ATP and risk assessment to identify factors for systematic study.

Typical uses of DOE in QbD are as follows:

1. Estimate effects of factors on responses
2. Study interactions between factors and their effects on responses
3. Estimate the precision of a measurement
4. Identify factors that have a significant effect on responses
5. Select optimum operating conditions and/or ranges
6. Identify factors that have little effect on responses (robustness studies)
7. Identify regions of failure
8. Reduce the number of factors (screening studies)
9. Reduce the number of factor levels
10. Identify factor ranges
11. Build empirical models to predict responses over the experimental range (response surfaces)
12. Estimate coefficients of known models

A general strategy for applying DOE to QbD is to perform a screening design (e.g. Plackett/Burman) to reduce the number of levels and/or factors so that a second study (e.g. fractional factorial) can be performed to further investigate the more important factors and to evaluate any possible interactions between the factors. Then, if desired, the final step is to use a response surface design so that an empirical model can be fit to establish a relationship between each response and the factors.

A DOE is used to evaluate the relationship between the factors and the responses. There can be multiple responses and/or factors in an experiment. In most cases the analysis consists of evaluating the effect of the factors on each response separately. This type of analysis is called a multifactor analysis. If the analysis is evaluating the effect of the factors on multiple responses in the same analysis (e.g. several different impurities), then the analysis would be considered multivariate. The advantage of a multivariate analysis is that it takes into account the correlations between responses. For example, several impurities may be related to each other—when impurity A is high, impurity B may tend to be low. The multivariate analysis would take this into account. At the present time, multivariate analysis is not commonly used for QbD since the analysis is much more complicated than multifactor analysis. Multivariate analysis using principal components or partial least squares is commonly used in building chemometric models.[11,12] Since multivariate analysis is beyond the scope of this chapter, only multifactor experiments will be discussed in the remainder of this section.

3.5.2 Multifactor designs

There are many types of multifactor designs that could be used in a QbD strategy, such as Full and Fractional Factorials, Nested, Split Plot, Mixture, and Response Surface designs. A number of texts have been written on the design and analysis of experiments,[13–18] which describe these designs in detail.

A design commonly used in the pharmaceutical industry consists of only one factor at several levels. For example, the OFAT strategy would be to perform a one-way experiment for one factor holding the other factors constant, then pick another factor holding the other factors (including the first factor) constant. For example, in a chromatographic experiment, all settings may be held constant except for flow rate. Flow rate could be set at specific values and several runs performed at each of these flow rate settings. This would be called a one-way experiment since there is only one factor (flow rate). This may be repeated, changing another factor while holding the rest constant. In the event of interactions between factors (see below), this is *not* an optimal strategy for multifactor experiments.

3.5.2.1 Factorial designs

Factorial designs are the most common designs used in QbD. These are used to identify important factors as well as any interactions that may exist between factors (see Chapter 5 in Ref. 13). These designs are used for method development as well as for showing the ruggedness or robustness of a method over a region. They consist of two or more factors with each factor set at two or more levels. The total number of combinations that could be tested is the product of all the levels. Each combination of factors and levels is called a treatment combination. So if there are two factors at two levels and one factor at four levels, there would be 2 × 2 × 4 = 16 treatment combinations. The design before randomization would look like the following (Table 3.2):

Table 3.2 Three-factor design (two at two levels, low, L, and high, H, and one at four levels, L1-4) prior to randomization

Run	Factor A (Two Levels: L and H)	Factor B (Two Levels: L and H)	Factor C (Four Levels: L1, L2, L3, L4)
1	L	L	L1
2	L	L	L2
3	L	L	L3
4	L	L	L4
5	L	H	L1
6	L	H	L2
7	L	H	L3
8	L	H	L4
9	H	L	L1
10	H	L	L2
11	H	L	L3
12	H	L	L4
13	H	H	L1
14	H	H	L2
15	H	H	L3
16	H	H	L4

It is important to run the 16 experiments in a random order to eliminate any systematic errors. Performing a full factorial design allows estimation of the effects that each factor has on the response as well as the possible interactions between the factors. In the example above, there are three factors, so the analysis would include estimation of the main effects, 2-way and 3-way interactions. Main effects of a factor are computed by determining the difference between the average of one level of the factor averaged over all the other factors to the average of another level of the factor averaged over all the other factors. So the main effect of A is the average of the responses corresponding to runs 1–8 to the computed average response from runs 9–16. The 2-way interaction between factors A and B would compare the four combinations of the A and B levels as shown in Table 3.3A. An example of an AB interaction is shown in Table 3.3B and Fig. 3.2.

In this example, the effect of factor A depends on the level of factor B. At the low level of factor B, increasing factor A from low to high increases the response by 13 but at the high level of B, increasing factor A from low to high increases the response by only 0.5. If a factor is involved in an interaction, then interpreting the factor's main effect can be very misleading. Notice that if an OFAT strategy was performed and the scientist held factor B at the high level first and performed a run at the low and high level of factor A, there would be no difference since they both would result in a measured value around 82. If the scientist followed up holding the factor A at the low level and performed a run at the low and high level of B, then the low level of B would indicate a lower response than the high level of B. The scientist would decide that the high level of B is optimum and factor A has little effect, completely missing the fact that the low level of factor B and high level of factor A would result in a response of over 90.

In running a factorial design, replication of points is highly recommended. In the above example, suppose the experimenter only obtained one result from each of the four combinations of factors A and B of 78.3, 91.1, 80.8, and 81.3 as shown in the table. It is possible that these four results could have been obtained by performing the same combination of factors A and B four times. The variability in the

Table 3.3A Two-way interactions are determined by comparing the average AB levels

		Factor A	
		L	**H**
Factor B	**L**	Average runs (1–4)	Average runs (9–12)
	H	Average runs (5–8)	Average runs (13–16)

Table 3.3B An example of the AB interaction

		Factor A	
		L	**H**
Factor B	**L**	78.3	91.1
	H	80.8	81.3

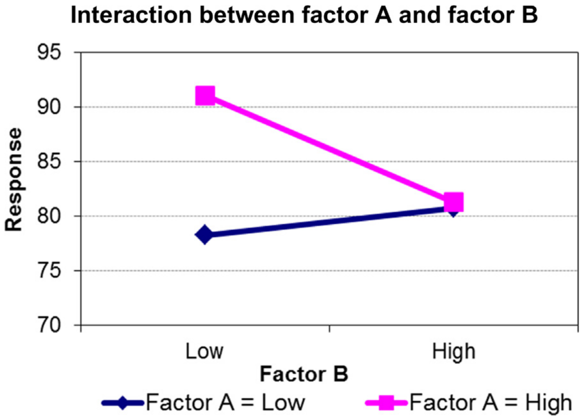

FIGURE 3.2

Graphical representation of the AB interaction shown in Table 3.3B. (For color version of this figure, the reader is referred to the online version of this book.)

results may not be due to changing the factor levels but rather just the natural variation in the method upon repeating the same treatment combinations four times. If the factors are all quantitative, then replication can be accomplished in a factorial design by adding center points at the mean of each quantitative factor. If the design also contains qualitative factors, then there is no "true" center. For example, if the design consists of the factors time (quantitative) at 5 and 10 min and two solvents (qualitative), then the replicates would be performed at 7.5 min for each solvent. Replication is used to test the significance of factor effects on the response and to provide an estimate of the reproducibility of the treatment combination. Another benefit of center points is that if the factors are quantitative and the center is the average of the high and low levels, then it is possible to obtain an estimate of curvature over the experimental region. If there is a significant curvature, then predicting the response within the factor ranges cannot be done accurately because the design cannot determine which factor is causing the curvature. Additional design points are needed to determine what factor(s) are causing the curvature. Response surface designs (Section 3.5.2.4) are often used to estimate curvature.

Performing a full factorial design with several factors each at several levels becomes large very quickly. For example, seven factors each at two levels would require 128 separate experiments! In this case, fractional factorial designs, which are subsets of full factorial designs, are generally used since they require fewer treatment combinations (see Chapter 6 in Ref. 13). These subsets are chosen in a special way so that the maximum information can be gained from the experiment. Fractional factorials generally provide less information on higher order interactions. For example, a full factorial design with five factors each at two levels would require $2^5 = 32$ treatment combinations. But a half fraction (expressed as 2^{5-1}) would require testing only 16 treatment combinations as shown in Table 3.4. The 2 represents the number of levels, the 5 represents the number of factors, and the -1 represents the fraction of the full factorial (a -2 would mean a quarter fraction).

There is some loss of information because the entire 32 run design is not performed. This is called confounding, meaning that certain terms are not separable from each other. In this design, the loss of information arises from the fact that the main effects are confounded with the four-way interactions

Table 3.4 2^{5-1} Fractional factorial design

Run	Factors				
	A	B	C	D	E
1	L	L	L	L	H
2	L	L	L	H	L
3	L	L	H	L	L
4	L	L	H	H	H
5	L	H	L	L	L
6	L	H	L	H	H
7	L	H	H	L	H
8	L	H	H	H	L
9	H	L	L	L	L
10	H	L	L	H	H
11	H	L	H	L	H
12	H	L	H	H	L
13	H	H	L	L	H
14	H	H	L	H	L
15	H	H	H	L	L
16	H	H	H	H	H

and the two-way interactions are confounded with the three-way interactions. For example, the interaction between A and B is confounded with the three-way interaction of C and D and E. If the AB interaction is significant in the analysis, the experimenter will not know whether the AB interaction is causing significance or the CDE interaction because they are indistinguishable. If the experimenter believes that the only possible effects are the main effects and the two-way interactions, and that the three- and four-way interactions do not exist or are very small, then not much is lost by running the ½ fraction.

3.5.2.2 Nested designs

These are often used to partition the total method variability into its contributing parts. For example, in an assay method validation, one could make three preparations on each of two days from the same batch and perform two injections for each preparation. The injections are nested in preparation and the preparations are nested in day. The design with results is shown in Table 3.5A.

The analysis of this design would separate the total variability into three parts (called components): between day, between preparations within day, and between injections within preparation. The analysis would provide the separation of variability as shown in Table 3.5B. As can be seen from the table, most of the variation is due to day-to-day variation. In QbD, this type of design and analysis can help to identify where the greatest sources of variability lie so that the experimenter knows where to put efforts to improve the method.

There is often confusion as to whether a design is a factorial or a nested design. For example, suppose that there are two factors, “method” and “batch”. If there are only three batches, A, B, and C,

Table 3.5A Example of a nested design, with three preparations on each of two days from the same batch, with two injections for each preparation

Day	Preparation	Injection number	Results
1	1	1	40.2
		2	41.8
	2	1	43.9
		2	44.2
	3	1	39.9
		2	38.8
2	1	1	43.4
		2	45.5
	2	1	46.0
		2	47.2
	3	1	46.2
		2	46.8

Table 3.5B Variability assigned to different factors

Source	Standard Deviation	RSD(%)
Day	2.90	6.64
Preparation within day	1.78	4.07
Injections within preparation	0.92	2.10
Total	3.52	8.07

and each batch is tested by both methods, then the design is a factorial design. However, if the batches A, B, C tested by method A are totally different from the batches D, E, F tested with method B, then the batches are nested in method. In the nested design, there are six batches but in the factorial design, there are only three batches.

3.5.2.3 Split-plot design

Suppose the first four runs after randomization in a design with factors A and B are as shown in Table 3.6. The experimenter notices that factor A is at the low level for runs 1 and 2. Therefore instead of resetting A to low again, the experimenter just leaves the setting alone. In an experiment, each run should be performed as though it is the first run. All levels should be reset. However, there may be practical reasons why it is difficult to reset the factor (sometimes called “hard to change” factor). For example, the experimenter may be studying different mobile phases, flow rates, and temperatures. Remaking the mobile phase for each run may not be easy so the experimenter may want to use one

Table 3.6 The first four runs after randomization in a design with factors A and B

Run	Factor A (Two Levels: L and H)	Factor B (Two Levels: L and H)
1	L	L
2	L	H
3	H	L
4	H	H

preparation of mobile phase for several combinations of flow rate and temperature before switching mobile phase. To the analyst, this seems perfectly reasonable; they are interested in the question of whether mobile phases A or B affect their separation, so why remake the mobile phase each time? On the other hand, a statistician would consider the preparation of the mobile phase to contribute its own variability to the method and would accordingly analyze the data differently, using a split-plot design (hence, the importance of analyst and statistician discussing how the experiment is designed before it is performed!). A split-plot design for this example (before randomization) would be as shown in Table 3.7.

The randomization occurs in two steps for a split-plot design since the four flow rate by solvent combinations have to occur in each of the four mobile phase preparations. In the example, the four

Table 3.7 Split-Plot design (before randomization)

Main Plot	Mobile Phase	Split Plot	Flow Rate	Temperature
1	A	1	L	L
		2	L	H
		3	H	L
		4	H	H
2	B	1	L	L
		2	L	H
		3	H	L
		4	H	H
3	A	1	L	L
		2	L	H
		3	H	L
		4	H	H
4	B	1	L	L
		2	L	H
		3	H	L
		4	H	H

mobile phase preparations would be randomized first. Then within each of the four mobile phase preparations, the four flow rate by solvent combinations would be randomized. If the first preparation of mobile phase A is first after randomization, then all four combinations of flow rate and solvent would be run before switching to the second mobile-phase preparation in the randomization. Split-plot designs were commonly used in agricultural experiments where one factor was used for a large plot (called main plot) of land and then other factors were applied to subplots of the main plot.[18] In the example above, the mobile phase is the main plot and the four combinations of flow rate and solvent are applied to portions of the same mobile phase and are called the split plots. If this experiment was a factorial design, then mobile phase would have to be prepared 16 times whereas in the split-plot design, mobile phase is only prepared four times (2 A's and 2 B's). The analysis of a split plot takes into account that mobile phase was only applied four times but flow rate by solvent combinations were applied 16 times. The analysis allows for two sources of variability—one for the main plots and one for the split plots. Main plot variability is used to evaluate the main-plot factors and the split-plot variability is used to evaluate the split-plot factors as well as the interactions between the main-plot and split-plot factors. Split-plot designs are discussed in detail in Chapters 10 and 11 of Ref. 19.

3.5.2.4 Response surface

The designs discussed above are generally used to estimate the "true" mean response at the specific combinations in the study rather than interpolate or extrapolate outside the ranges used in the study. Response surfaces are very helpful in determining what factors are important, what the effect of changing the factor levels have on the response, estimating experimental error, and evaluating interactions between factors. However, if only two levels are used for quantitative factors, then interpolation or extrapolation can be very risky since a linear relationship must be assumed. Adding center points can allow an estimation of the overall curvature but cannot identify which factor is causing the curvature. In order to interpolate or extrapolate, designs should use an adequate number of levels to allow a reliable prediction equation. The designs discussed above can be used for this purpose by adding additional factor levels.

Response surface designs are used to develop a function that will relate the responses to the factor levels. The factors are generally quantitative. These designs help the experimenter to visualize the effects of the factors on the response. There are several texts on response surface designs.[20,21]

The empirical model that is generally used for response surface designs is a full quadratic model that includes the linear, cross product, and quadratic terms. An example of a full quadratic model in two factors is as follows:

$$\text{Response} = B_0 + B_1{}^*X_1 + B_2{}^*X_2 + B_{12}{}^*X_1{}^*X_2 + B_{11}{}^*X_1^2 + B_{22}{}^*X_2^2 \tag{3.3}$$

where

B_0 = intercept
B_1, B_2 = linear term coefficients
B_{12} = cross product coefficient
B_{11}, B_{22} = quadratic coefficients.

Since the "true" relationship between the response and the factors may be complicated, these models are approximations to the "true" model (see Chapter 10 in Ref. 13). Therefore, the model should be evaluated as part of the analysis to ensure that the model is fitting the data adequately. The range over which the model is used is also important. A full quadratic model may be adequate for a limited region of experimental space even if it cannot be used for a larger region.

Least squares are used to fit a model to the data with no assumptions to fit the model. However, to make any statistical statements such as determining significant terms, constructing confidence intervals, or lack of fit, the following assumptions are made:

1. The model is correct. There are several ways that a model can be incorrect. One incorrect model is overfitting the data. For example, a full quadratic model could be fit to a response but the true model just contains the linear term. Alternatively, two factors may be included in the model but only one factor has an effect on the response. The other incorrect model is underfitting the data. In this case, there are not enough terms in the model. An example of this is fitting a line to show linearity of reference response when the true model is quadratic. One approach is to fit the quadratic model and test for curvature by testing that the coefficient associated with quadratic term is significant. R^2 is not a good measure of linearity since the R^2 measures the combination of curvature as well as random variation about the regression line.
2. The observations about the fitted model are independent of one another.
3. The underlying distribution of the residuals about the fitted line is normal.
4. Variability of the residuals is similar at each combination of factors in the experiment. If this assumption is not satisfied, there are several strategies to correct the problem. For example, the data could be transformed so the residuals are normal or a weighted regression could be used. Weighted regression requires knowledge about the relationship between the factor settings and the variability or enough data at each treatment combination so that an estimate of the variability at each treatment combination can be made. This variability can be used as weights in the weighted regression.

A factorial design that has three levels for each factor can be used to fit a full quadratic model. The design for a 3^2 factorial (two factors each at three levels for nine treatment combinations) for the factors X_1 and X_2 is shown in Fig. 3.3.

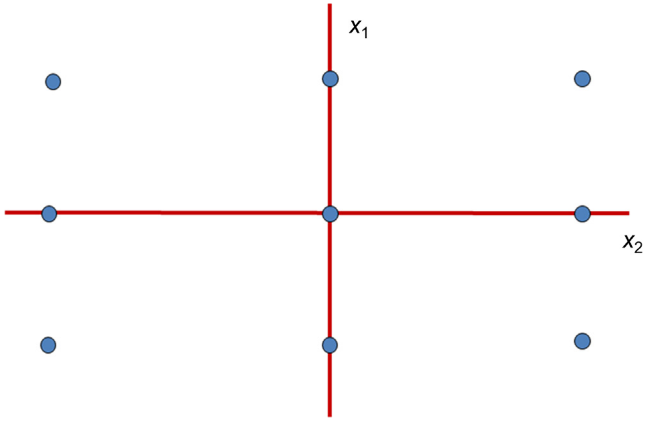

FIGURE 3.3

Three level factorial. (For color version of this figure, the reader is referred to the online version of this book.)

Table 3.8A An example of a 3^2 factorial to evaluate the effect of concentration and time on a response (recovery) before randomization

Concentration	Time	Response
1	10	73
1	20	85
1	30	80
3	10	80
3	20	90
3	30	84
5	10	85
5	20	96
5	30	90

An example of a 3^2 factorial is a study to evaluate the effect of concentration and time on a response such as recovery. The design before randomization is given in Table 3.8A. The least squares full quadratic model fit to the data provides estimates of the coefficients and the *p*-value associated with each term (Table 3.8B).

In Table 3.8B, the *p*-value is the probability of finding a coefficient as large as or larger than found in the study assuming that the "true" coefficient is zero. *p*-values less than 0.05 are considered significant since they would occur rarely if the "true" coefficient was zero. Therefore the conclusion is that the "true" coefficient is not zero. There is a significant quadratic effect of time but not concentration. Both time and concentration are involved with significant effects. In an attempt to simplify the model, one approach would be to eliminate the concentration-squared term since it is not significant. However, this approach should be performed with some caution since the design is small with low power to find significant effects. At a minimum, the fit of the full model against the actual results with the model not including the concentration-squared term should be examined to evaluate the effect of removing the term. The response surface and contour plots based on the full model are shown in Fig. 3.4. The surface plot shows the curvature in time and a rising predicted response as concentration increases. If the goal is to maximize the response, then the optimum is on the edge of the experimental region with time at around 20 and concentration of 5. However, estimating the response at given

Table 3.8B 3 × 3 factorial quadratic model coefficients

Term	Estimate	*p*-Value
Intercept	42.18	<0.01
Concentration	3.00	0.04
Time	3.68	<0.01
Concentration × concentration	0.042	0.76
Concentration × time	−0.025	0.25
Time × time	−0.083	<0.01

combinations of time and concentration is difficult to do visually from the response surface plot (Fig. 3.4(a)); the contour plot (Fig. 3.4(b)) is much better for this purpose. Any combination of time and concentration on the same contour line predicts the same response.

The number of runs required to conduct a full 3^k factorial can be problematic. Two factors require nine runs, three requires 27, and for four factors, 81 runs are necessary. Even with three factors, the number of runs may be too large in practice. However, there are other designs that allow fitting a quadratic model and do not require as many runs. The most common design is the central composite (CCD), which consists of a 2-level factorial design, center points, and axial (or "star") points. For three factors, the design is shown in Fig. 3.5.

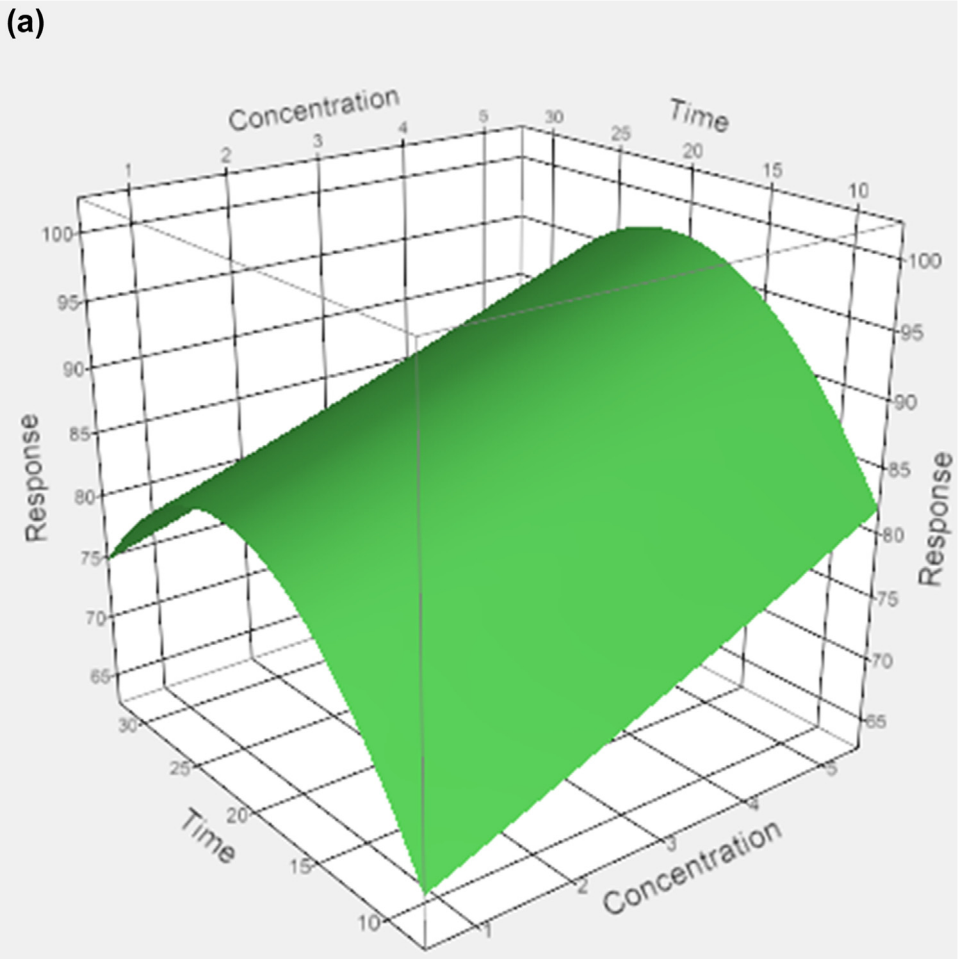

FIGURE 3.4

(a) Response surface. (b) Contour plot. (For color version of this figure, the reader is referred to the online version of this book.)

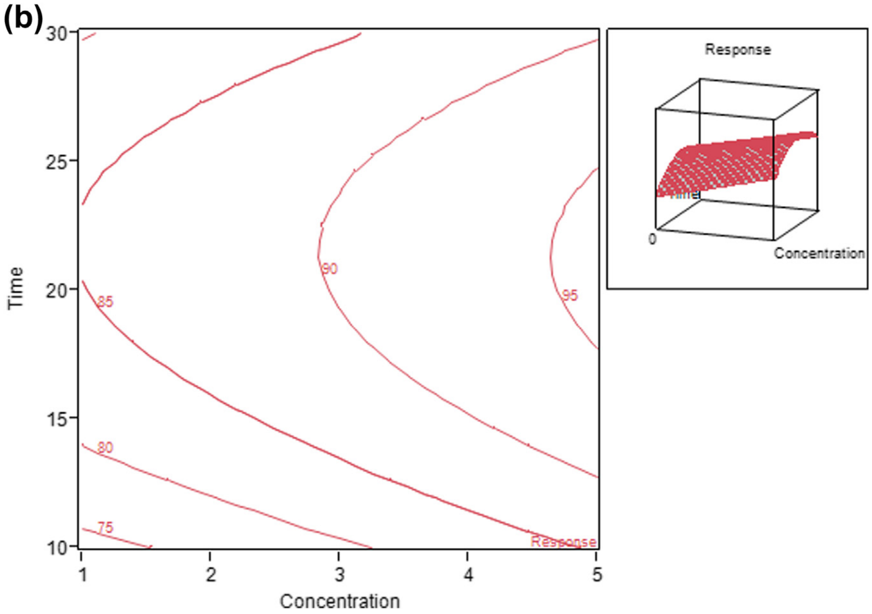

FIGURE 3.4

(*continued*).

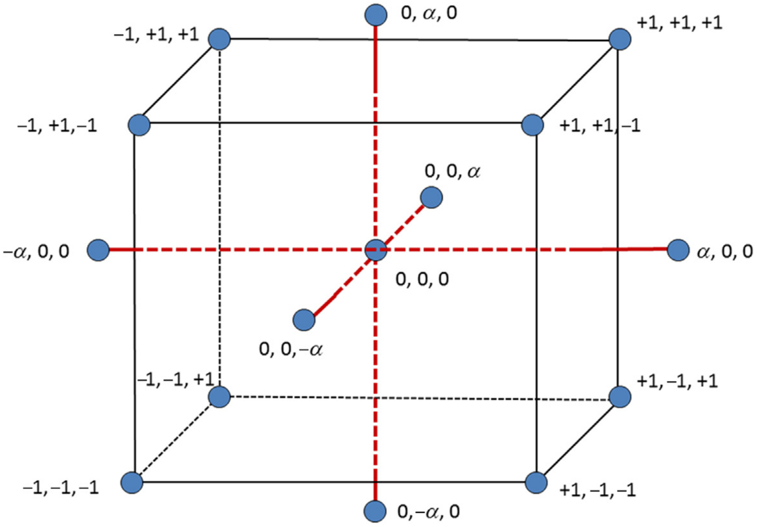

FIGURE 3.5

Central composite design for three factors. The corners of the box are the factorial portion of the design, with additional center points (0,0,0) and axial points (points with α in their coordinates). (For color version of this figure, the reader is referred to the online version of this book.)

Table 3.9 Number of runs (factorial + star + center + additional centers) required to run a central composite design

# Factors	Central Composite
k	$2^k + 2k + 1 + \text{AC}$
2	9 + AC
3	15 + AC
4	25 + AC
5	27 + AC

Table 3.9 shows the number of runs (factorial + star + center + additional centers (AC)) required to run a central composite design. Note that the number of points in the central composite is much smaller than in a 3^k factorial design.

Placement of the star points and/or number of center points can give the design different properties. The distance from the factorial points on each axis to the center point is considered one unit. The multiple of this distance is usually denoted by α. So if $\alpha = 1$, the star points would align with the factorial points. In the case of two factors, this would result in a 3 × 3 factorial. An $\alpha = 1.414$ would result in a design that is called rotatable, meaning that for any concentric circle about the center, any predicted response on the circle would have the same precision. Other values are possible, resulting in different properties.

A case study, which illustrates the use of a response surface design in sample extraction is discussed in Section 3.8.

3.5.2.4 Mixture designs

Mixture designs are used when the factor levels are proportions of a total amount. For example, a solution such as a mobile phase or an extraction solvent may consist of three components with each component representing a percentage of the total, for example 20% component A, 30% component B, and 50% component C. The sum of the proportions adds up to 100%. The goal of the study may be to find the optimum combination of factor percentages. For example, the experimenter may be interested in finding the best combination of surfactant, solvent, and oil to increase recovery. The design can be described in the diagram shown in Figure 3.6, and is also given in Table 3.10. This particular mixture design is called a lattice with each component ranging from 0 to 1 by thirds. Points are denoted by (water, surfactant, oil) giving the proportion of each component in the mixture (Fig. 3.6, Table 3.10). Note that a factorial or central composite design (CCD) cannot be used in these studies since the sum of the factor levels in some treatment combinations could add up to more than 100%. For example, if component A is to be studied between 10% and 30%, B between 20% and 40%, and C between 40% and 60%, a factorial would require a combination with A = 30, B = 40, and C = 60%, which is impossible since the components add to 130%!

Note that the interaction terms are not significant. One could investigate whether or not to reduce the model by eliminating these terms (one at a time). The corresponding contour plot using the full model is given in Fig. 3.7, showing the combinations of water and surfactant that predict the value on the contour line. The oil content would be 1 minus the sum of the water and surfactant levels.

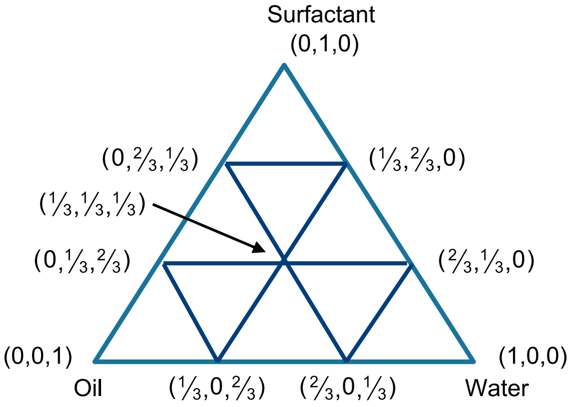

FIGURE 3.6

Lattice mixture design for three components. (For color version of this figure, the reader is referred to the online version of this book.)

Table 3.10A Input values and responses for the mixture design shown in Fig. 3.6

Water	Surfactant	Oil	Response
1	0	0	70.8
1/3	0	2/3	79.7
2/3	0	1/3	81.2
0	0	1	78.7
1/3	1/3	1/3	68.2
2/3	1/3	0	60.7
0	1/3	2/3	68.7
1/3	2/3	0	75.2
0	2/3	1/3	79.3
0	1	0	70.3

Table 3.10B The resulting model

Term	Estimate	*p*-Value
Water	69.913	0.0005
Surfactant	75.007	0.0003
Oil	76.776	0.0003
Water × Surfactant	−24.733	0.50
Water × Oil	27.802	0.40
Surfactant × Oil	−12.890	0.69

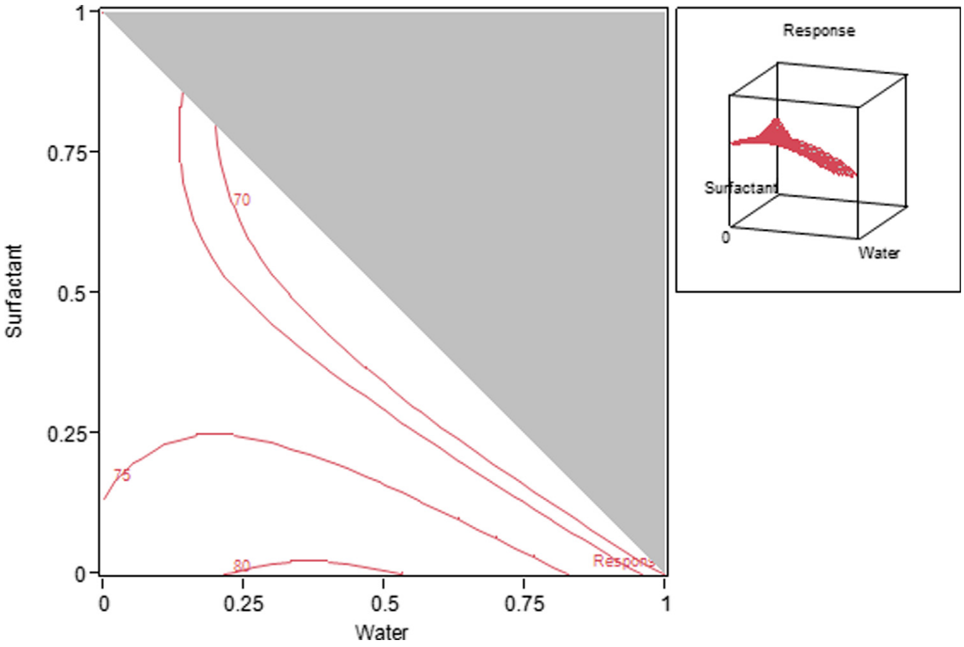

FIGURE 3.7

Contour plot for the mixture design data in Table 3.10. (For color version of this figure, the reader is referred to the online version of this book.)

To maximize the response (greater than 80) would require a surfactant level less than 5%, a water level between 25% and 50%, and depending on the choices of surfactant and water, an oil level between 45% and 75%.

3.5.2.5 Optimal designs

Prior to selecting a design, the scientist should think about the goals of the experiment. In addition to the goal, the experimental region of interest should be selected. In many experiments such as those described above, the region is rectangular by default since each single factor has a range and when all factors are combined, the result is a multidimensional rectangle. However, there may be problems where the experimental region of interest is not rectangular due to physical constraints. One example is the mixture design described above where the experimental region is triangular or a prism. If two factors are flow rate and temperature, there may be a different range of usable flow rates depending on the temperature. Another question to ask prior to creating a design is whether or not the goal is to find the best model by selecting the important main effects, interactions, or quadratic effects or to assume a specific known model but try to find the best estimates of the coefficients (or obtain the most precise predicted response). Optimal designs[19] are often used in cases where the experimental region is non-rectangular but can be defined and/or when the model can be specified but the goal is to obtain the "best" model. The "best" model could mean finding the model with the most precise coefficients or the most precise predicted value or many other statistical criteria. The most common criteria when the goal is to obtain the most precise estimates of the coefficients in the model is D-optimality, whereas the most common criteria to obtain the most precise predicted value over the experimental region are either G- or I-optimality[19].

Once the experimental region is defined (which can be all of the possible combinations or factor levels in the experiment), the model specified, number of runs are chosen and optimality criteria selected, then the optimal design needs to decide what treatment combinations to run and how many replications to perform at each treatment combination. There are various software packages such as JMP that will generate an optimal design. If the goal of the experiment is to find the best model, then the scientist should be careful when performing the analysis since the coefficients are correlated (partially confounded) with one another. Therefore, the coefficient for a term in the equation depends on whether or not other terms are included in the model. The correlation also affects the significance of terms in the model. Optimal designs can result in a large reduction in the number of runs to perform in an experiment.

3.6 APPROACHES USING EXPLICIT MODELS

The previous section focused on using empirical models in QbD. However, it may be possible to build an accurate mechanistic model, which describes the system being studied. The application of empirical and mechanistic modeling to a QbD study of chemical reactions has been compared[22]; the authors found both approaches could describe their reaction. Building the mechanistic model had the advantage of developing greater process understanding, greater ability to explore transient conditions, and aided in risk assessment by the ability to rapidly conduct simulations to test the sensitivity of the reaction to various factors. On the other hand, the empirical model offers an approach where an adequate mechanistic model cannot be developed, e.g. due to the complexity of the system. For analytical applications, several commercial software packages are available for chromatographic method development and optimization that are based around well-established models of chromatographic retention.[23] Arguably, the best known of these is DryLab, which has been available in increasingly sophisticated versions for a quarter of a century.[24] Chromatographic retention is modeled using a variety of expressions which describe the effects on retention of parameters such as mobile-phase composition, temperature, pH and additive concentration[25]:

$$\text{Solvent strength (\%B): } \log k = \log k_{\mathrm{w}} - S\varphi \tag{3.4}$$

where k is the retention factor of the analyte in the aqueous–organic mobile phase (k being the ratio of the amount of analyte in the stationary phase to that in the mobile phase, a value which can be related to the retention time of the analyte), k_{w} is the retention factor of the analyte using a mobile phase comprised only of water, S is the solvent strength parameter for this analyte, and φ is the volume fraction of organic solvent in the mobile phase.

$$\text{Temperature: } \log k = A + B/T \tag{3.5}$$

where A and B are constants for a given system, and T is the temperature (in K); this is essentially a simplified expression of the Van't Hoff equation.

$$\text{Mobile phase pH: } k = k^{0}(1 - F) + k^{\mathrm{i}}F \tag{3.6}$$

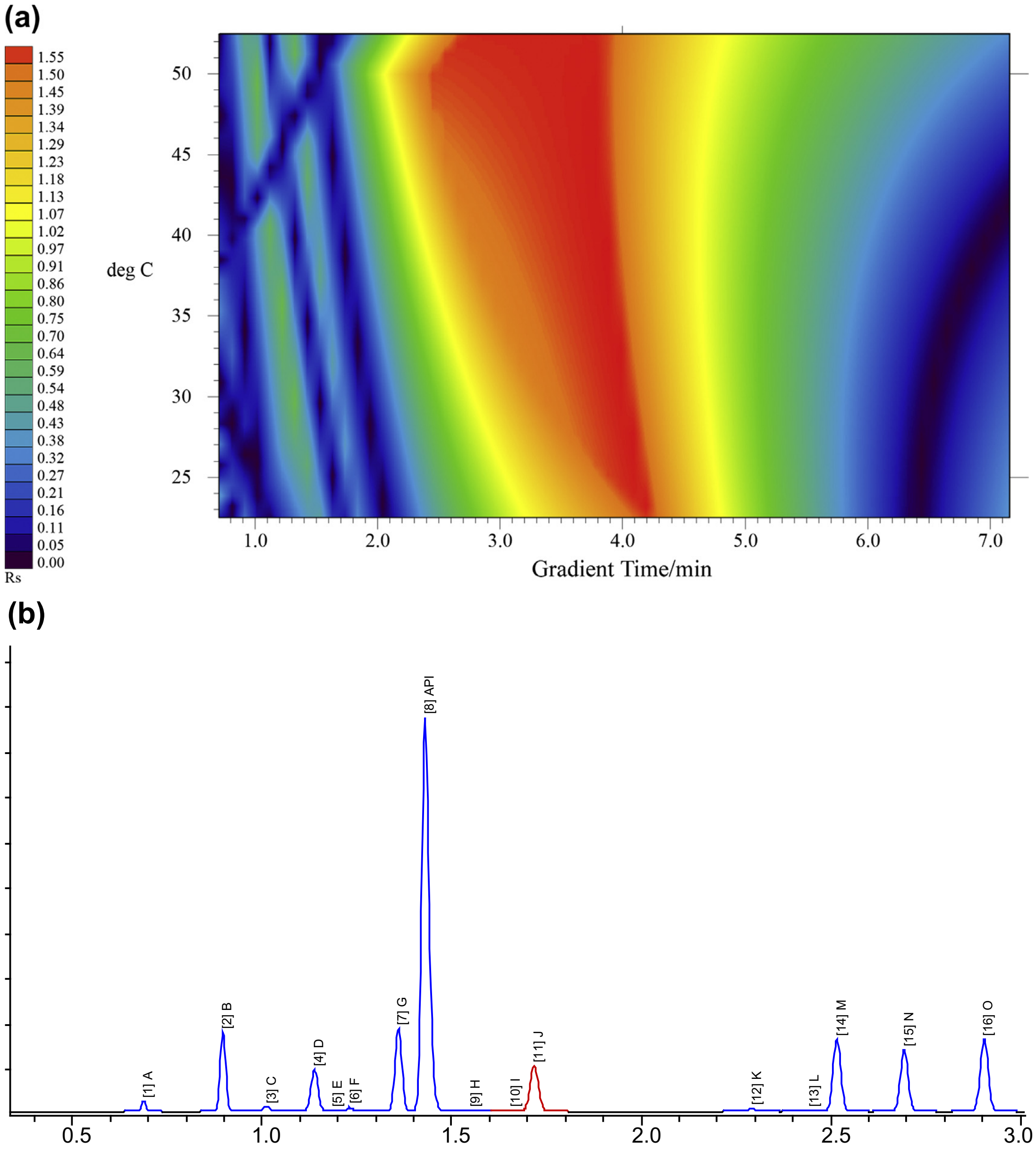

FIGURE 3.8

(a) Resolution map generated using Drylab. The scale indicates values of R_s achieved under the various separation conditions used. Optimum resolution occurs at higher temperature over a range of gradient times from approx. 2.5–4.0 min. (b) Predicted chromatogram at a gradient time of 2.5 min, at 50 °C. Peaks I (very small impurity, not visible on this scale) and J are the critical pair under these conditions. (c) Actual chromatogram. Only main components are identified with retention times. (For color version of this figure, the reader is referred to the online version of this book.)

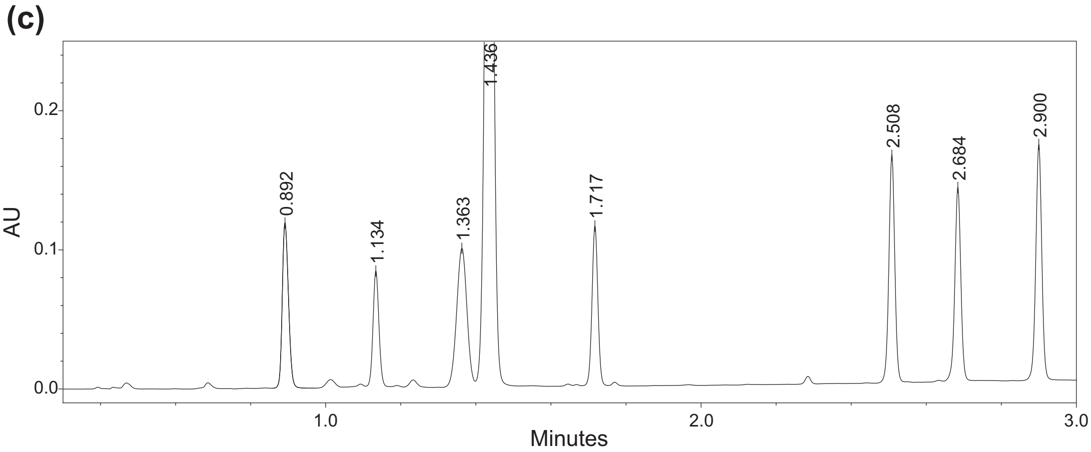

FIGURE 3.8

(*continued*).

where k^0 is the retention factor of the analyte in its neutral form, while k^i is the retention factor of the ionized species, and F is the fraction of the analyte that is ionized, which can be determined via the Henderson–Hasselbach equation (Eqn (3.2)).

$$\text{Buffer/additive concentration: } \log k \approx C + D \log [X] \tag{3.7}$$

where C and D are constants for a given system, and $[X]$ is the concentration of the interacting additive such as an ion-pairing agent.

Overall retention is determined as a combination of the effects of the individual parameters. Although these are a mix of empirical and more fundamental expressions, they are well established as reasonably accurate descriptions of the effects of key chromatographic variables on retention. The values of the coefficients are determined in a small number of experiments; not all variables need to be studied in each case, e.g. if there are no additives or ionizable solutes, these additive concentration and pH effects are not studied. Thus, a useful model can be obtained *with a very limited number of input experiments* (a considerable practical advantage for explicit models, when available). Although it is understood that there are some interactions when multiple parameters are changed, the effects are quite limited and, generally, accurate predictions are achieved.[25,26] As well as retention, peak width and shape are modeled, and the output is a resolution map, illustrating critical resolution between peaks as a function of parameters such as analysis time and temperature, for a given combination of solvent, column dimensions, flow, etc. An example is shown in Fig. 3.8(a), illustrating a map of resolution for a gradient very high pressure liquid chromatographic separation (resolution, R_s, is a measure of the separation of two peaks in a chromatogram based on the width of the peaks and their separation; $R_s > 1$ and preferably >1.5). The input data were from just four chromatographic runs, at two temperatures and two gradient times. From this limited input data the method was optimized. In the response surface (Fig. 3.8(a)), warmer colors indicate greater resolution between the critical pair, and an optimum region can be seen to exist in the central region of the map, with gradient times of around 2.5–4 min and temperatures in the range 40–50 °C. The predicted and actual chromatograms are shown in Fig. 3.8(b and c), illustrating the high degree of accuracy that is achieved (predicted and actual retention times agree within 2 s in this example). It should be noted that although the 2-D map in Fig. 3.8(a) plots resolution as a function of column temperature and gradient time, the effect of variables such as column length and diameter, flow rate, and gradient profile can be determined from the same data since these are accounted for in the underlying models (other parameters such as pH may also be varied if these are part of the model used and appropriate data are collected). Recent versions of DryLab are particularly focused on QbD applications, and offer 3-D visualization of the design space.[27,28]

3.7 GENERAL ADVICE ON DESIGN/ANALYSIS OF EXPERIMENTS

This section contains advice on the design and analysis strategy of experiments. As Eleanor Roosevelt said, “Learn from the mistakes of others. You can’t live long enough to make them all yourself.”

3.7.1 Design strategy

1. Talk to other scientists: if you are performing your first designed experiment, talk to other scientists who have already completed a design. They can provide valuable information on setting up the equipment, obtaining appropriate materials, problems encountered in setting up and running the experiment, collecting the data, formatting the data for analysis, and lessons learned.
2. Ask whether the design will answer the right question: be sure of the question before designing the experiment; think about the question that you are trying to answer. You don’t want to complete the experiment, analyze the data and find out that it is not addressing the right question. One strategy once the design is determined is to enter simulated data using values which are realistic for the proposed experiment. Then analyze it and review the results.
3. Include relevant players: prior to designing an experiment, think about the scientists who will be affected by the conclusions. Include all relevant players in planning the study. The method is often transferred to another department that may have constraints that do not allow the method to be run in the same way as was optimized. If you have access to statisticians, do not wait until the data are collected before getting them involved. Most statisticians are trained to design experiments as well as analyze them.
4. Pick meaningful factor levels: after performing the risk analysis, most factors will be determined for study in the designed experiment. However, one must still pick the levels for each factor. This can be the hardest part of designing the experiment. If the levels are too close together, it will be difficult to find any effects while if the levels are too far apart, it is possible that a large number of treatment combinations will fail to provide meaningful results. It is also possible that the underlying relationship between the response and the factors has “cliffs” or nonlinear areas that are not fit well by the statistical model. An example would be an acid–base titration; this could be modeled by a linear or quadratic equation over a short range, but not for a wide range of factor levels. On the other hand, a good fit over a wide range could be made from limited data if the correct equation describing the titration process was used. In the early stages of method

development, it is desirable to allow the levels to be more spread out, but when finalizing the method for robustness or to develop a "design space", the levels should be picked over a range that provides flexibility and keeps the responses within specifications or internal limits.

5. Record data to the appropriate number of significant figures: data should be recorded with enough digits as to make the analysis reliable. Suppose a degradant is the measurement and all of the data are recorded to one place past the decimal point ranging from 0.5 to 0.8. This causes the precision estimate to be inaccurate, which in turn makes the analysis less accurate in terms of significance. It should be noted that the ICH Q3A and B Guidelines on Impurities in the Drug Substance and Drug Product, respectively, describe the number of significant figures that should be used in *reported* data. This is not meant to imply that these significant figures are appropriate for calculation of secondary data.
6. Record results and observations: it is important to keep detailed written notes during the performance of the study. This can be very helpful when the analysis shows some "outliers" or unusual results. Knowing that something different occurred during that particular run may explain the problem. This can also be useful when transferring a method. Sometimes a little change in technique may not be captured in the method (e.g. the way a vessel is shaken, or the position of a flask within an ultrasonic bath).
7. Replicate: as noted in the previous section, replication is an important part of a designed experiment. Usually the replicates are performed at the center but can also be obtained by replicating the design. However, replicating the design can result in expending greater resources. Suppose that an experiment has two factors each at two levels. Then replication could be accomplished by running a 2^2 factorial with four center points or by replicating the whole 2^2. The advantage of replicating the center points is that a measure of curvature can be obtained. Another advantage is that since the center point may be the desired settings for the method, additional data at this point may be helpful. If the whole 2^2 is run, then one gains additional precision information on each factorial point precision. The greater advantage is that the effects are estimated more accurately since they are averages of replicates. Adding centers does not have this property since center points do not increase the number of results at each factorial point.
8. Perform pilot runs: it is possible to run DOE too early in the development process. The designed experiment is not the place to still be learning how to run the equipment or learning the basics of the method. Also, pilot runs are useful to help establish levels for the factors. One strategy in picking levels is to run the "worst" case prior to starting the designed experiment. This may not be an easy decision because the "worst" case may not be all factors high or all factors low. So the decision of "worst" case prior to performing many runs may need to be based on the science.
9. Consider running designs in sequence: during the development process, it is common to perform more than one design. Based on the analysis of the first design, a scientist may decide to run a second design that may use the same factors but different factor levels or may add/eliminate factors. Planning of the second design should include thought as to what factors and levels were used in the first design. It is common that each design is analyzed completely separately. However, if the second design is well planned, the designs can be combined into a single analysis that provides much more information. One example is called the fold-over design. Suppose a design contains four factors each at two levels in a half fraction of a 2^4 factorial. This design uses eight of the possible 16 treatment combinations. The design confounds two-way interactions with each other. So if a two-way interaction is significant, one cannot tell

which of the two confounded two-way interactions is affecting the responses. A second design could be run by using the eight treatment combinations that were left out in the first design. This is a fold-over and allows all two-way interactions to be estimated, thus eliminating the confounding problem. Another example is using central composite designs. As discussed previously, the central composite design consists of a factorial, center points, and axial points. Instead of running the entire design before analyzing, one could run just the factorial part and some center points. If the results indicate that the factor levels were chosen so that the design is in the area of interest, then the second design would include the axial points as well as additional center points. When running designs in sequence, you should always use common points in both designs (usually the center) so that you can detect if a shift has occurred from the first design to the second design. This could indicate that something has changed and an investigation may be needed.

10. Consider blocking: blocking can be very useful to evaluate factor effects. Blocking is done by grouping the treatment combinations within a homogeneous set. For example, suppose that an experiment consists of two factors, A at two levels and B at three levels for a total of six treatment combinations. Each combination is used to prepare tablets that will be tested for dissolution. Since the dissolution apparatus usually consists of six vessels, there are two ways to perform the dissolution testing: (1) For each of the six treatment combinations, test six tablets with all six of the same treatment combination tested in the same dissolution apparatus or (2) test one tablet from each of the six treatment combinations in the same dissolution apparatus (the block) and run each set six times. The total number of tablets tested is 36 for each possibility but the second method is much better since all six treatment combinations are tested within the same dissolution apparatus making a much better comparison of the two factors since the apparatus run to run variation is eliminated. Another example would be comparing two potency assays using multiple batches (the block) of tablets. Instead of using one potency method on some batches and the other method on other batches, both assays would be used on each batch. Then the difference between the two assays has lower variability since the batch-to-batch variability has been removed. Blocking is discussed in Chapter 2 of Ref. 13.

3.7.2 Analysis strategy

Once the results of the study are available, a recommended strategy for analysis is as follows:

1. Review raw results: look for extreme (unexpected) observations or entry mistakes.
2. Review center points if available: since the center points were all performed at the same combination of factor levels, they should reflect the reproducibility of the factor combination. If this result is much higher than expected, this may indicate a problem with the experiment. The center point variability is used for testing the effects. If the variability is very low, then smaller differences between factor levels and interactions are more likely to be found significant. Similarly, high variability would require larger differences in the effects to be found significant. If the variability is higher than expected, it is possible that there are other sources of variation that are not being accounted for in the study.
3. Evaluate assumptions: as stated above, certain assumptions are made when analyzing data from an experiment. If these assumptions are not satisfied, then the *p*-values, which indicate significant effects, are affected. In many of the experiments described above that were not response surface

or mixture designs, it is often the case that only one result is available for each treatment combination, so checking assumptions can be difficult. Randomization is important to obtain independence of the results. Checking for normality and equal variance is also difficult to do in these situations. However, with response surface designs, there are plots that can help to check assumptions, such as a plot of the residuals against predicted results or against each factor. The plots should not show any patterns in the residuals of the factorial experiments described in the previous section.

4. Examine highest order interactions first: in most experiments, main effects and two-way interactions are of most interest. In this case, the two-way interactions should be examined first.
5. Examine main effects/interactions not involved in higher order interactions: as stated in List 4, if the experiment only contains main effects and two-way interactions, then main effects that are not involved in a two-way interaction should be examined. The reason for this is that if a main effect is involved in a two-way interaction, then the effect of that factor depends on the level of another factor.
6. Examine results of the curvature test.

3.7.3 Finding the best operating point

The goal of many experiments is to find the best combination of factors to either maximize (e.g. recovery) or minimize (e.g. impurity) the response or find the combination closest to a target value (e.g. label claim). The experimenter runs a DOE and obtains the responses. One option (not the best one) is to find the best result among the responses in the experiment. The results could be sorted from high to low and then just chose the "best" one. Then the combination of factors associated with that response is chosen at the optimum condition. The better option is to analyze the data either by estimating effects or fitting a response surface and determining which effects are significant based on the *p*-value.

Factorial designs use statistical significance to find the "BEST". If the experiment used a fractional factorial design, then not all combinations of factors were used in the experiment (e.g. a half fraction of a factor design would only use 16 of the possible 32 treatment combinations). Therefore, the best combination may be one of the treatment combinations that were not run in the experiment. The statistical analysis can be used to find the best combination even though it was not in the experiment.

3.7.4 Causes of nonstatistical significance

After running an experiment, it can be frustrating when the found effects are not significant. For example the main effect of a factor on potency is 8% but was not significant. This is usually due to the study not having enough power to detect the difference. As part of the planning for an experiment, the number of replicates of the center and treatment combinations should be considered so that there is an assurance that if a meaningful difference really exists, the design will find the difference significant in the analysis. It is dangerous to make decisive conclusions on effects that were not significant. The statistical analysis determines if the difference could have happened by chance. So if the effect is not significant, then it is possible that there is no difference and making a decision based on this inconclusive result could result in a bad decision. High variation in the center points is a sign that small differences will not be considered significant. Adding more than 4 or 5 center points loses the ability to find significant differences. Instead of adding center points, additional factorial points should be added to the experiment. This results in a better estimation of the effects since the number of points used to calculate means is increased. Another cause of nonsignificance is outliers. This could be due to a high-order interaction or an error that occurred during the experiment. Another possibility when no significant effects are found is that the factors have no effect within the experimental region. The goal of robustness studies may be to show no effects over the region so no significant effects can be a desirable result (as long as the study was large enough to detect significant effects).

3.8 CASE STUDY—SAMPLE EXTRACTION METHOD DEVELOPMENT USING A RESPONSE SURFACE DESIGN

3.8.1 Problem statement

A potency and impurities assay was being developed for a solid oral dosage form. Extraction was required prior to chromatographic analysis.

The compound is a small molecule which is poorly soluble and not very stable in water. It is highly soluble in a variety of organic solvents, and is more stable in aprotic solvents such as acetonitrile. The experimental formulation consisted of small sugar beads coated with active drug and a protective polymer to prevent drug degradation in the acidic stomach. These enteric-coated beads were then filled into a capsule.

3.8.2 Analytical target profile

The extraction objectives, and desired performance characteristics, are shown in Table 3.11.

Initially, an attempt was made to develop manual sample preparation methods; however, significant degradation was observed as the drug was exposed to water during this procedure. The drug is relatively stable in acetonitrile, but since the protective polymer coating is not soluble in acetonitrile, directly placing the intact beads in this solvent was not an option. As an alternative, beads were manually ground into a powder, followed by extraction of the drug from the powder using acetonitrile. However, this procedure was lengthy, irreproducible, and raised significant safety concerns in handling of this highly potent compound. Therefore, an alternative approach was required, and automation was chosen, using a Tablet Processing Workstation II (TPWII).

Table 3.11 Analytical target profile for extraction method

Extraction objectives	• Complete and reproducible extraction of drug from the capsules • Minimal degradation during the extraction process • Safe process
Performance criterion 1	Method accuracy better than ±2%
Performance criterion 2	Method precision better than ±2%
Performance criterion 3	Degradation during sample preparation $<0.1\%$
Performance criterion 4	Minimize analyst exposure to drug

3.8.3 Extraction method description

The TPWII has a robotic arm for transfer of sample capsules from input test tubes to a vessel containing a homogenizer probe. Solvent is added to the vessel and the dosage forms under test are broken up by the homogenizer in a series of pulses where each pulse is a period of time in which the homogenizer probe spins rapidly. The vessel is made to cycle up and down during part of the homogenization to ensure that capsules are drawn up into the homogenizer blades. A "mixing time" step was added, during which there is a low-speed rotation of the homogenizer probe. Under these low-energy conditions, the capsules are not broken up but the mixture in the vessel is gently stirred. This step allows extra time for the drug to dissolve after the beads have been broken up in the initial vigorous homogenization. Part of the vessel contents is then pumped through a filter and into a high-performance liquid chromatography sample vial.

3.8.4 Risk analysis

Since the TPWII removes the majority of the analyst's contact with the samples, the automated extraction approach effectively addresses the ATP's performance criterion 4—safety. Therefore, the risk analysis focused on the extraction performance in terms of criteria 1–3. Several operating parameters can significantly affect the sample preparation process, including solvent type, solvent volume, speed of homogenization, number and duration of homogenization pulses, mixing time, filter type, flush volumes used to rinse the instrument tubing, and vessel washing parameters. Potential critical parameters were identified in a brainstorming session, guided by existing knowledge of the product and of the automation platform used. A fishbone diagram illustrating the risks considered is shown in Fig. 3.9. The risks are grouped according to different parts of the TPWII sample preparation process:

Homogenization: It is intuitively obvious that the homogenization speed (measured in rotations per minute of the homogenizer probe) and the time for which homogenization takes place will

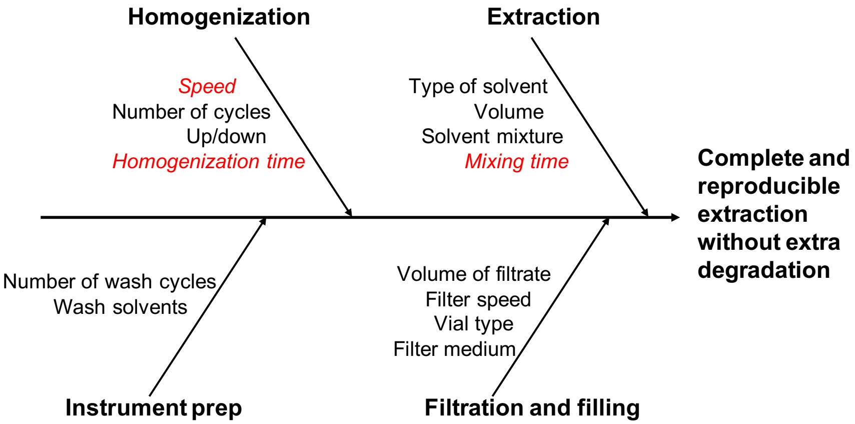

FIGURE 3.9

Risk assessment for the automated extraction method. (For color version of this figure, the reader is referred to the online version of this book.)

affect the disruption of the capsules. In addition, it was known from preliminary experiments that with more vigorous homogenization, increased degradation could occur. Thus, homogenization time and speed were considered critical for further study. The number of homogenization pulses was considered less critical—the total time being considered more important than the way the homogenization is delivered. Similarly, the number of up/down cycles may affect whether all capsules actually engage the blades and break up; clearly an intact capsule would lead to lack of extraction, but by observation after a few cycles all capsules were broken and so this was not considered a critical parameter for systematic investigation.

Extraction: As well as the physical elements of homogenization, the chemical process of drug dissolution had to be considered. Based on prior knowledge of the analyte, its stability was inadequate in water or alcohols, but good in acetonitrile. Solubility was also very high in acetonitrile. Thus the extraction solvent was fixed as acetonitrile without further study, and the volume was not considered critical because of the high analyte solubility. However, the mixing time was chosen for further study, since it was reasonable to believe that the length of time that the drug was in solution in contact with the excipients may affect its stability.

Instrument preparation was considered. The TPWII flow paths can be washed with solvents before and after use. Because of the known lability of the analyte to water, only pure acetonitrile was chosen as wash solvent; by procedurally eliminating water from the system this factor was adequately controlled and further study was not needed.

Filtration/filling: Based on prior knowledge of the compound, the filter and vial type were not considered critical. Factors such as the volume of filtrate and filter speed were not considered likely to interact with other experimental parameters, and were thus optimized separately in a univariate fashion.

3.8.5 Experimental design

A two-level factorial design was initially performed, which confirmed both the significance of the three factors chosen in the risk analysis, and that there were interactions between them. A CCD was then used to generate response surfaces for measured potency and degradation as a function of the factors homogenization time, homogenization speed and mixing time (t_h, s_h, and t_m). The CCD consisted of 15 points (2^3 = eight full factorial points, one center point, and six star points). Lower and upper limits for the three factors used in the factorial part of the design were $t_h = 100$ and 600 s, $s_h = 12{,}000$ and 18,000 rpm, and $t_m = 0$ and 300 s. The center point of the design was at $t_h = 360$ s, $s_h = 15{,}000$ rpm, and $t_m = 150$ s. The star points were chosen due to instrumental constraints: $t_h = 60$ and 850 s, $s_h = 10{,}000$ and 20,000 rpm, and $t_m = 0$ and 400 s. In addition to the three factors described above, batch-to-batch and day-to-day effects were also evaluated by running the 15 point CCD each day on several days for two different batches. One batch was tested on three days while the other batch was tested on one day. Details of the design and the associated potency and degradation results are presented in Table 3.12.

Note that the potencies are consistently higher for batch B, and the impurities are lower when compared to batch A; this reflects the actual characteristics of the batches rather than any effect of the extraction process. The data for potency and degradation were fit to each batch separately using JMP software to a full quadratic regression model that included linear, quadratic, and cross

Table 3.12 Experimental design and measured data. One batch was run on three days, a second batch on one day. The order of experiments was randomized on each day. Experimental conditions and potency and degradant results are listed in sorted factor order

Data Display										
			Potency				Degradant			
			Batch A			Batch B	Batch A			Batch B
Homogenization Time/s	**Homogenization Speed/1000 rpm**	**Mixing Time/s**	**Day 1**	**Day 2**	**Day 3**	**Day 1**	**Day 1**	**Day 2**	**Day 3**	**Day 1**
60	**15**	**150**	94.74	96.27	94.18	96.01	0.27	0.28	0.25	0.12
100	**12**	**0**	87.70	88.91	89.51	84.85	0.24	0.27	0.28	0.12
100	**12**	**300**	97.39	97.43	98.95	100.83	0.26	0.27	0.27	0.12
100	**18**	**0**	92.05	94.11	94.31	95.29	0.26	0.26	0.25	0.12
100	**18**	**300**	98.32	100.19	97.86	101.12	0.27	0.27	0.28	0.11
360	**10**	**150**	99.17	98.06	98.47	102.62	0.25	0.29	0.27	0.14
360	**15**	**0**	99.44	99.51	99.13	103.38	0.28	0.28	0.28	0.12
360	**15**	**150**	99.36	97.12	97.93	103.67	0.28	0.27	0.29	0.13
360	**15**	**400**	99.57	98.78	98.14	104.10	0.30	0.29	0.30	0.14
360	**20**	**150**	99.20	100.11	96.90	100.89	0.33	0.33	0.31	0.17
600	**12**	**0**	99.45	98.98	99.06	102.64	0.28	0.29	0.29	0.13
600	**12**	**300**	99.62	98.75	98.90	102.18	0.29	0.28	0.32	0.16
600	**18**	**0**	98.48	99.03	99.14	101.85	0.39	0.34	0.38	0.21
600	**18**	**300**	99.29	98.61	97.81	101.96	0.43	0.44	0.47	0.30
850	**15**	**150**	98.03	98.87	99.30	100.21	0.41	0.46	0.39	0.26

product terms for the quantitative continuous factors t_h, s_h, and t_m. The qualitative factor day was included in the model for batch A to estimate the day-to-day variation and determine whether or not day was a significant factor in the model. The estimated day-to-day standard deviation for the potency was 0.90% (relative to the label claim) with a p-value of 0.76 (not significant). Similarly, for the amount of degradant, the estimated day-to-day standard deviation was 0.017% degradant, with a p-value of 0.66 (not significant). Therefore, the day term was eliminated from the model. The remaining quadratic model was then reduced by eliminating terms that were not significant (p-values ≥ 0.10) starting with the quadratic terms, followed by the cross product terms, and finally linear terms. If a higher order term was significant, then any lower order term contained in that factor was kept in the model. For example, if mixing time squared was significant, then the linear mixing time term was kept no matter whether it was significant or not since it is a factor in the squared term.

The final models for potency and degradation are shown in Tables 3.13 and 3.14.

Since there was replication for batch A, an F-test was performed to test if the model adequately fit the potency and degradant data (lack of fit), which indicated that there was no significant lack of fit for either response (p-values ≥ 0.12).

It can be seen that the same factors can be used in models describing both the measured potency and degradation.

Table 3.13 Batch A: Factors included in the final model for potency and degradant, estimates of their coefficients and significance of each term

Batch A				
Parameter Estimates				
	Potency		Degradant	
Term	Estimate	*p*-Value	Estimate	*p*-Value
Intercept	78.65	<0.0001	0.397	<0.0001
Homogenization time/s	0.0451	<0.0001	−0.0006	<0.0001
Homogenization speed/ 1000 rpm	0.6841	0.0002	−0.0091	0.0009
Mixing time/s	0.0529	<0.0001	−0.0003	0.0612
(Homogenization time/s)^2	−1.828e−5	<0.0001	2.1499e−7	0.0003
Homogenization time/s × Homogenization speed/ 1000 rpm	−0.0011	0.0031	3.8576e−5	<0.0001
Homogenization time/s × Mixing time	−0.00005	<0.0001	2.2009e−7	0.0435
Homogenization speed/ 1000 rpm × Mixing time	−0.0012	0.0508	2.2222e−5	0.0157
(Mixing time)^2	−2.939e−5	0.0291	−1.269e−7	0.5262

Table 3.14 Batch B: Factors included in the model for potency and degradation and significance of each term

Batch B				
Parameter Estimates				
	Potency		**Degradant**	
Term	**Estimate**	***p*-Value**	**Estimate**	***p*-Value**
Intercept	73.45	<0.0001	0.231	0.0061
Homogenization time/s	0.0810	0.0068	−0.0006	0.0068
Homogenization speed/ 1000 rpm	0.8739	0.1302	−0.0070	0.1154
Mixing time/s	0.0397	0.0055	−7.97e−5	0.3537
(Homogenization time/s)^2	−3.768e−5	0.0187	2.0361e−7	0.0728
Homogenization time/s × Homogenization speed/ 1000 rpm	−0.0020	0.1598*	3.7931e−5	0.0050
Homogenization time/s × Mixing time	−7.456e−5	0.0200	4.2999e−7	0.0615

Although this p-value was greater than 0.10, the term was kept in the model to keep the models consistent for plotting purposes. This term for potency could be deleted and the model fit again, if desired.

In Fig. 3.10, contour plots are shown which illustrate the response surfaces for potency and degradation obtained as a function of t_h and s_h at t_m values of 0, 150 and 300 s. Any point on the same contour has the same predicted potency (or degradation). For example, in Fig. 3.10(a), any combination of homogenization time and speed associated with the blue line labeled 0.28 has a predicted degradant level of 0.28%. A homogenization time of 500 s with a homogenization speed of 13,000 rpm or a homogenization time of 300 s with a homogenization speed of 17,800 rpm have a predicted degradant level of 0.28%.

It can be seen from Fig. 3.10(a), when $t_m = 0$, maximum extraction is only achieved at large values of t_h. This is somewhat improved by increasing s_h only when t_h is relatively low. On the other hand, the minimum of degradation only occurs at $t_h < 300$ s. Increasing t_m to 150 s brings the optimum regions closer together, with a clear plateau for potency seen at lower values of t_h in Fig. 3.10(b). This trend continues as t_m is increased to 300 s, and in Fig. 3.10(c), the region of maximum potency is seen to closely approach the area of minimum degradation. A global optimum of method performance exists at the intersection of the individual optimum regions of the contour profiles for recovery and degradation. From Fig. 3.10(c), it can be seen that optimum conditions are approximately $s_h = 12{,}000$ rpm, $t_h = 400$ s and $t_m = 300$ s. Although this does not correspond to the absolute minimum for degradation, a difference of 0.01% is not significant and so these extraction parameters represent a good compromise. Greater t_h or s_h would place the method on the rapidly rising part of the degradation surface, which is not considered acceptable.

The forms of the response surfaces are in agreement with the interpretation that more vigorous conditions (longer extraction, higher homogenization speed) lead to more complete extraction of the

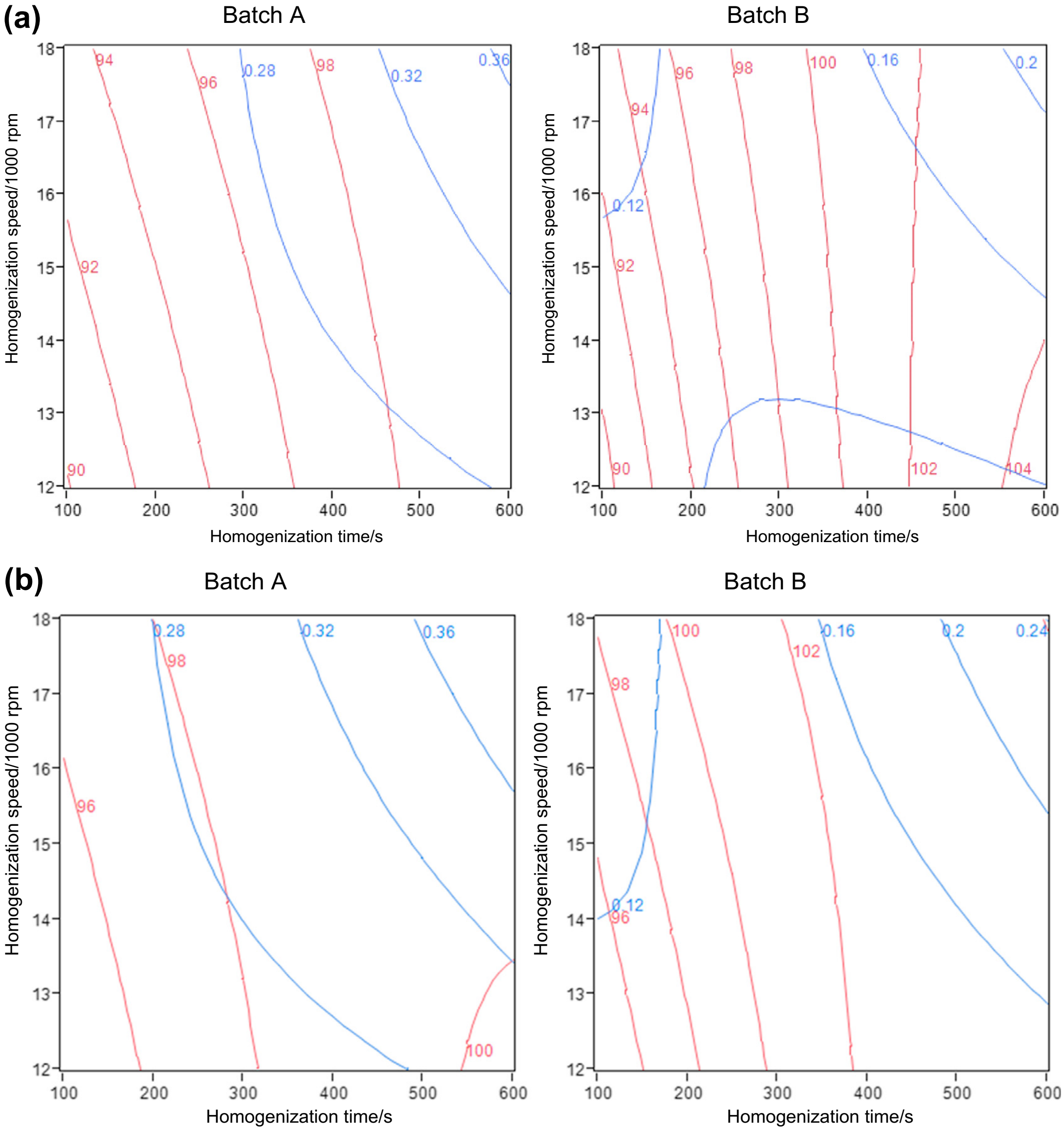

FIGURE 3.10

Contour plots for potency and degradant as a function of t_h and s_h at different values of t_m. The optimum region exists at the point where potency is maximized, and amount of degradant is minimized. (a) $t_m = 0$ s, (b) $t_m = 150$ s, (c) $t_m = 300$ s. (For color version of this figure, the reader is referred to the online version of this book.)

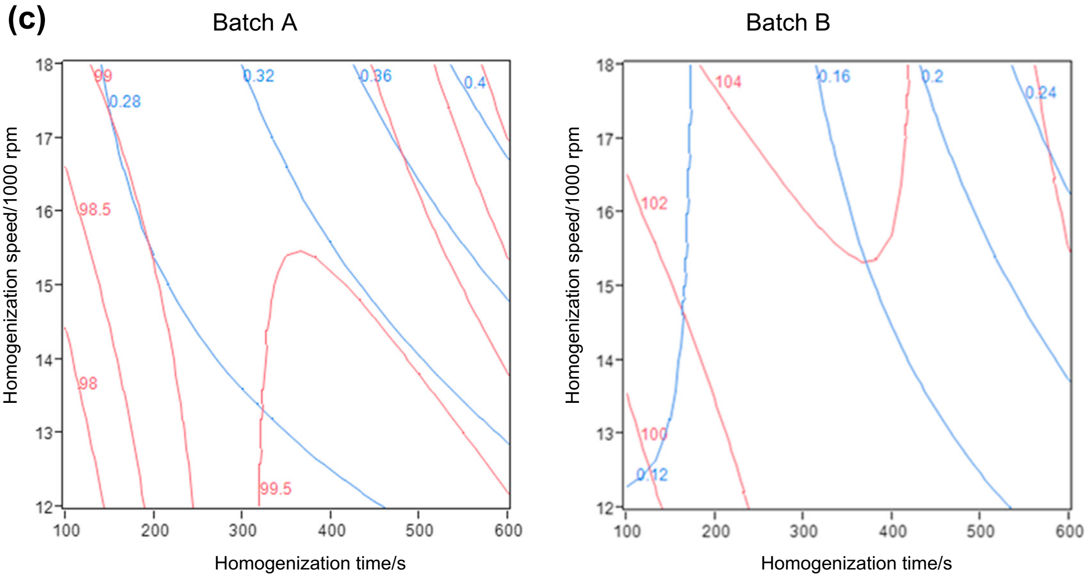

FIGURE 3.10

(*continued*).

drug from the capsules, whilst promoting degradation. It is interesting to note that the potency response surface begins to curve down at the most vigorous extraction conditions, e.g. dropping below 99% for $t_h > 500$ s and $s_h > 16{,}000$ rpm (Fig. 3.10(c)). This may to some degree reflect true changes in the amount of drug in solution, since the measured potency will decrease as the amount of degradation increases. However, with a predicted potency loss of greater than 1% under the most vigorous extraction conditions, the increase in degradant is only around 0.2%. Since the response factors of both compounds are similar, this could be interpreted as predicting a mass balance deficit. However, this apparent deficit is due to the limitations of the quadratic fit in modeling the sigmoidal relationship between the extraction conditions and the measured potency; the model describes the response surface as being a symmetrical hill, when in reality it is more like an asymmetrical plateau.

3.9 DEVELOPMENT TO VALIDATION

Regulatory guidance exists on the required elements of validation (see Ref. 29 and Chapter 4 of this book), and validation is typically performed as a discrete activity at the end of the development process. This guidance has proved extremely useful in standardizing expectations for method validation, but a consequence has been that validation tends to follow a rigid, procedure-driven path for determination of accuracy, precision, etc. Following a QbD approach to method development, the systematic studies performed should result in extensive knowledge of the primary factors which are critical to successful method operation, demonstrated operating ranges, and data on method performance within those ranges. Appropriate controls will also have been identified, either as elements of the method itself, or as qualification requirements for instruments, SOPs for operators or facilities, etc. Therefore, it can be argued that in a QbD approach a final validation in its traditional format is not required; much of the method performance is defined during method design (in studies which are scientifically justified for the method under study, but which may differ greatly from method to method). Consequently, it has been proposed that in analytical QbD a life cycle approach be adopted, comprising method design, method qualification (involving a modest degree of experimentation to demonstrate the method meets the requirements laid out in the ATP under routine operating conditions; perhaps such studies can simply be documented from the development phase), and by continued verification of method performance during the method lifetime.[30] It will be interesting to see how validation guidance evolves in the future to incorporate such concepts.

ICH guidelines[29] include some well-defined approaches for validation of a variety of method performance parameters. However, for assessment of method robustness the guidance is less specific: “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.” More generally, a method should be rugged, i.e. insensitive to factors external to the method, such as where it is run and by whom. A QbD approach to method development facilitates achieving these goals. The risk assessment process identifies primary factors (those expected to have a significant effect on the experiment), and the modeling process demonstrates the range within which adequate performance is achieved as these key parameters are varied. So, in the chromatographic example described above, the response surface in Fig. 3.8(a) illustrates the sensitivity of the method to changes in gradient time and temperature. Similar maps may be generated after making small variations in the mobile phase composition, flow rate, etc., to determine whether the method is sensitive to these factors. Thus, a region may be defined where adequate resolution is achieved for any operating parameter setting. Within this, the method operating space may be defined (likely to be smaller than the absolute maximum ranges determined from the model). Running the method at the extremes of the range defined can verify the predicted performance. In the case of a chromatographic method, where resolution is typically modeled, verification runs may be useful in that other attributes such as accuracy and sensitivity may be checked. If a factorial or response surface design has been performed with replication, this will demonstrate the method performance across the studied space.

Although considerable understanding is gained during method development, the more sophisticated models created will typically include a small subset of possible method parameters (primary factors) for thorough investigation, as defined via risk analysis. Secondary factors identified during risk assessment (those not expected to have a significant effect on the method) will likely not be extensively studied during method development. However, secondary factors should still be evaluated to make sure that they do not have an unanticipated effect. So should both primary and secondary factors be studied in a screening study first in a highly fractionated design as part of risk analysis followed by selection of the most important factors for follow up experiments so that interactions can be studied? Or should the primary factors be evaluated first holding the secondary factors constant, then after optimizing the responses, perform a second study showing that the secondary factors originally left out have no or little effect on the responses? The advantage of the first approach is that one finds out early if any of the factors that were not expected to have an effect really do have an effect. If this happens, then follow up experiments can include those factors. The advantage of the second approach is that the first step may result in changing the factor levels to optimize the method. Then the secondary factors can be evaluated against the optimum factor levels. However, if there are secondary factors that have a significant effect on the response or interact with the factors already studied, then additional work would be required. The first approach may be better as long as there are not too many factors and the design does not require too many runs. On the other hand, if there is a high confidence that the risk analysis really has identified key parameters (e.g. through extensive knowledge of the technique employed) the second approach may be justified. A separate ruggedness study at the end of the process may in any case be required to encompass a broader range of factors that were not known or available when the original method was developed, e.g. to test the method against formulations prepared at the limits of the product design space, or to include testing at multiple sites. To limit the amount of work involved, very sparse designs may be employed for robustness studies.[31,32] If a factor is identified as important at this stage, then it will be necessary to add further controls and/or redefine the method operating space to ensure robust operation.

3.10 KNOWLEDGE MANAGEMENT

If QbD involves developing a full understanding of how method attributes and operating conditions relate to method performance, there needs to be a suitable mechanism for gathering all this knowledge together throughout the life of the method (indeed, knowledge management is an expectation outlined in ICH guidance[4]). A fundamental first step is to ensure that method development experiments are adequately documented with the reasons for performing the experiment and a conclusion based on the results gained. Exercises such as risk assessment or choice of study design should also be appropriately documented, such that the rationale for the decisions made is not lost. The approach taken to systematically collect the knowledge gained will depend greatly on questions such as the infrastructure available within a given organization. For example:

- Paper-based records, e.g. paper lab notebooks. Indexing and retrieval of data from QbD experiments is a considerable challenge. This may be aided by generation of a contemporaneous method development report, listing, for example, experiments performed and critical conclusions.
- Electronic notebooks (ELNs) offer much better search, indexing and retrieval capabilities than in the paper world. However, some systems are better than others and so it will likely be helpful if the analyst is systematic in using appropriate identifiers such as keywords so that method development records can be linked together. There is still an argument for creating an overview record which identifies key experiments and conclusions. Integration of data collection, modeling packages and corporate documentation systems with ELNs may allow a comprehensive solution to QbD data and knowledge management.
- Specialized data management systems have been proposed for analytical QbD. They may include the option to import data into shared, standard tools for analysis and report generation.

The final output may include method history, development and performance reports as well as the method description including validated operating ranges if desired. This package of knowledge becomes part of the transfer to receiving laboratories where the method will be put into use. Furthermore, it provides the basis for any future revisions.

3.11 QbD THROUGHOUT THE METHOD LIFETIME

Once the method is put into routine use at one or more quality control laboratories, a wealth of data will be generated which will indicate how it is performing, including:

- Simple observations by operators. Does the method continue to perform "as advertised" or are adjustments needed, e.g. a factor such as instrument equilibration time was identified as noncritical during development, but now extra equilibration time is needed.
- Is method performance changing, e.g. if there are system suitability criteria, is system suitability routinely met? Is there other evidence of the method misbehaving, e.g. out-of-specification (OOS) results (ones where the analysis is found to be the root cause, but also ones where the root cause is indeterminate and thus may be related to the analysis)?
- Systematic data collection and analysis, e.g. analyses of reference material data or system suitability data interpreted using control charts[33] to monitor method performance over time.

Observation of a pattern such as repeated OOS results related to the method, or a drift in quantitative performance is cause for further investigation and remediation. It should be rare that a new risk factor is identified at this stage, but this is not impossible. For example, an unannounced modification to the manufacture of a chromatographic column could result in a change which simply falls outside of the experimental space investigated in developmental studies, and may not become apparent until after aberrant results are generated and an investigation performed. Changes to site facilities and personnel may present similar challenges, although such changes are typically planned and thus can be prepared for. A change to the product which moves the product outside of the range of samples studied during method development may require the method to be reassessed, possibly even to the point of reevaluating the ATP. Such modifications to the analytical methodology should be planned in conjunction with the process modification, within the context of the firm's change management procedures.[4]

3.12 CONCLUSIONS

Although analytical QbD is not as well established as the application of QbD to product development, there is the potential for significant benefit in terms of robust performance of a method throughout its life. Definition of the ATP allows the method goals to be clearly stated, and risk analysis allows the development effort expended to be focused in the most important areas. Systematic studies allow the definition of a method operating space, extensive study of primary factors, and screening of a broader range of factors to determine robustness. If the comprehensive knowledge acquired during development becomes part of the package transferred to labs which will actually run the method, this will form the basis for understanding method performance and for continuous improvement.

The above description includes a variety of elements which, individually, are valuable. Hopefully, the whole is greater than the sum of the parts, and a revolutionary approach to the full implementation of QbD in the analytical laboratory could involve considerable upheaval, albeit with the maximum potential benefit. Alternatively, a step-by-step approach to implementing analytical QbD may be advocated, first incorporating elements of the analytical QbD toolkit where they make most sense in terms of existing workflows, and then looking for the greatest gaps in existing practices where most benefit can be gained. Within an organization, many individual elements of QbD may already be practiced, but perhaps in an informal way, or with less consistency than desirable. These are areas where gains can be made for a relatively modest effort, potentially acting as stepping-stones on the path to a more comprehensive application of analytical QbD.

Acknowledgments

The authors would like to acknowledge the work of Katie Wu who performed the chromatographic work described in Section 3.6 and the extraction experiments described in the case study in Section 3.8, and Emil Friedman who performed the initial analysis of the extraction data. The authors also acknowledge members of the QbD team in BMS Analytical and Bioanalytical Development for many discussions related to analytical QbD, in particular, Mario Hubert, Ruben Lozano, Adele Patterson, Yueer Shi, Liya Tang and Peter Tattersall.

References

1. *ICH Q8 (R2) – Guidance for Industry, Pharmaceutical Development,* International Conference on Harmonization, 2009.
2. McCurdy, V. Quality by Design. In *Process Understanding: For Scale Up and Manufacturing of Active Ingredients;* Houston, I., Ed.; Wiley, 2011.
3. *Quality Risk Management (Q9),* The International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, Second Revision 2005.
4. *Pharmaceutical Quality System (Q10(R4)),* The International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, 2009.
5. Borman, P.; Chatfield, M.; Nethercote, P.; Thompson, D.; Truman, K. The Application of Quality by Design to Analytical Methods. *Pharm. Technol.* **2007,** *31,* 142–152.
6. Torbeck, L.; Branning, R. QbD: Convincing the Skeptics. *BioPharm Int.* **2009,** *22,* 52–58.
7. Schweitzer, M.; Pohl, M.; Hanna-Brown, M.; Nethercote, P.; Borman, P.; Hansen, G.; Smith, K.; Larew, J.; Carolan, J.; Ermer, J.; Faulkner, P.; Finkler, C.; Gill, I.; Grosche, O.; Hoffmann, J.; Lenhart, A.; Rignall, A.; Sokoliess, T.; Wegener, G. Implications and Opportunities of Applying QbD Principles to Analytical Measurements. *Pharm. Technol.* **2010,** *34* (2), 52–59.
8. Nasr, M. Quality by Design (QbD): Analytical Aspects. In *32nd International Symposium on High Performance Liquid Phase Separations and Related Techniques;* Baltimore: MD, 2008.
9. Franklin, B. D.; Shebl, N. A.; Barber, N. Failure Mode and Effects Analysis: Too Little for Too Much? *BMJ Qual. Saf.* **2012,** *21,* 607–611.
10. Bowles, J. B. An Assessment of RPN Prioritization in a Failure Modes Effects and Criticality Analysis. *J. IEST* **2004,** *47,* 51–56.
11. Frank, I. E.; Friedman, J. H. A Statistical View of Some Chemometric Regression Tools. *Technometrics* **1993,** *35,* 109–148.
12. Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987,** *2,* 37–52.
13. Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters: Design, Innovation and Discovery,* 2nd ed., Wiley, 2005.
14. Montgomery, D. C. *Design and Analysis of Experiments;* Wiley, 1997.
15. Hicks, C. R.; Turner, A. V. *Fundamental Concepts in the Design of Experiments;* Oxford University Press, 1999.

16. Wu, C. F.; Hamada, M. *Experiments: Planning, Analysis and Parameter Design Optimization;* Wiley, 2000.
17. Cornell, J. A. *Experiments with Mixtures;* Wiley, 1990.
18. Milliken, G. A.; Johnson, D. E. *Analysis of Messy Data* In: *Designed Experiments,* Vol. 1; Chapman and Hall, 1992.
19. Goos, P.; Jones, B. *Optimal Design of Experiments: A Case Study Approach;* Wiley, 2011.
20. Myers, R. H.; Montgomery, D. C. *Response Surface Methodology;* Wiley, 2002.
21. Box, G. E. P.; Draper, N. R. *Empirical Model-Building and Response Surfaces;* Wiley, 1987.
22. Hallow, D. M.; Mudryk, B. M.; Braem, A. D.; Tabora, J. E.; Lyngberg, O. K.; Bergum, J. S.; Rossano, L. T.; Tummala, S. An Example of Utilizing Mechanistic and Empirical Modeling in Quality by Design. *J. Pharm. Innov.* **2010,** *5,* 193–203.
23. Garcıa-Alvarez-Coque, M. C.; Torres-Lapasio, J. R.; Baeza-Baeza, J. J. Models and Objective Functions for the Optimisation of Selectivity in Reversed-Phase Liquid Chromatography. *Anal. Chim. Acta* **2006,** *579,* 125–145.
24. Molnar, I. Computerized Design of Separation Strategies by Reversed-Phase Liquid Chromatography: Development of DryLab Software. *J. Chromatogr. A* **2002,** *965,* 175–194.
25. Dolan, J. W.; Lommen, D. C.; Snyder, L. R. High-Performance Liquid Chromatographic Computer Simulation Based on a Restricted Multi-Parameter Approach: I. Theory and Verification. *J. Chromatogr. A* **1990,** *535,* 55–74.
26. Snyder, L. R.; Dolan, J. W.; Lommen, D. C. High-Performance Liquid Chromatographic Computer Simulation Based on a Restricted Multi-parameter Approach: II. Applications. *J. Chromatogr. A* **1990,** *535,* 75–92.
27. Molnar, I.; Rieger, H. J.; Monks, K. E. Aspects of the “Design Space” in High Pressure Liquid Chromatography Method Development. *J. Chromatogr. A* **2010,** *1217,* 3193–3200.
28. Monks, K.; Molnar, I.; Rieger, H. J.; Bogati, B.; Szabo, E. Quality by Design: Multidimensional Exploration of the Design Space in High Performance Liquid Chromatography Method Development for Better Robustness Before Validation. *J. Chromatogr. A* **2012,** *1232,* 218–230.
29. *Validation of Analytical Procedures: Text and Methodology (Q2(R1)),* The International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, 1995.
30. Nethercote, P.; Ermer, J. Quality by Design for Analytical Methods: Implications for Method Validation and Transfer. *Pharm. Technol.* **2012,** *36,* 74–79.
31. Box, G. E. P.; Hunter, J. S.; Hunter, W. G. Additional Fractionals and Analysis. In *Statistics for Experimenters: Design, Innovation and Discovery;* Wiley: Hoboken, 2005; pp 281–316.
32. Torbeck, L. D. Ruggedness and Robustness with Designed Experiments. *Pharm. Technol.* **1996,** *21,* 169–172.
33. Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Control Charts, in Handbook of Chemometrics and Qualimetrics: Part A;* Elsevier: Amsterdam, 1997. pp 151–170 (chapter 7).