

DOE: The Jewel of Quality Engineering

11.1 Introduction

Design of experiments (DOE) methods are among the most complicated and useful of statistical quality control techniques. DOE methods can be an important part of a thorough system optimization, yielding definitive system design or redesign recommendations. These methods all involve the activities of experimental planning, conducting experiments, and fitting models to the outputs. An essential ingredient in applying DOE methods is the use of procedure called “randomization” which is defined at the end of this chapter. To preview, randomization involves making many experimental planning decisions using a random or unpatterned approach.

The purpose of this chapter is to preview the various DOE methods described in Part II of this book. All of these DOE methods involve changing key input variable (KIV) settings which are directly controllable (called factors) using carefully planned patterns, and then observing outputs (called responses). Also, this chapter describes the “**two-sample t-test**” method which permits **proof** that one level of a single factor results in a higher average response than another level of one factor. Two-sample t-testing is also used to illustrate randomization and its relationship with proof.

Section 2 provides an overview of the different types of DOE and related methods. Section 3 describes two-sample t-testing with examples and a discussion of randomization. Section 4 describes an activity called “randomization”, common to all DOE methods and technically required for achieving proof. Section 5 summarizes the material covered. Note that most of the design of experiments presented here are supported by standard software such as Minitab®, DesignExpert®, and Sagata® DOEToolSet and Sagata® Regression. (The author of this book is part owner of Sagata Ltd.; see www.sagata.com for more details.)

11.2 Design of Experiments Methods Overview

Five classes of experimentation and analysis methods are described in this book: (1) two-sample t-tests, (2) standard screening using **fractional factorials** (FF), (3) one-shot response surface methods (RSM), (4) sequential response surface methods, and (5) Robust Design based on Profit Maximization (RDPM). A brief summary is offered in Table 11.1. In addition, two classes of analysis of variance (ANOVA) analysis methods have been provided for determining significance after data has been collected using any experimental plan.

The primary objective is to allow the reader to develop competence in application of methods in each class. Also, decision support information for supporting has been provided for the selection of specific methods of each type, *e.g.*, choosing the number of runs, n , and the parameters used in the analysis. Note that any of these methods could constitute an entire “improvement system”.

Besides randomization, a common aspect of all DOE methods is the importance for the method users in identifying the KIVs and ranges for these factors. The preliminary identification of KIVs derives from engineering judgment. If a poor choice of KIVs and/or ranges is identified, it is unlikely the application of any DOE method will achieve desired results.

Note that all of the methods in Table 11.1 can generate statistical “proof” that changing factors affects average system outputs or responses. In general, derivation of the associated statistical proof relates to the amount and quality of the data collected and not whether the differences detected are important to decision-makers. An important theme in design of experiments is that statistical significance and evidence do not generally translate into “practical” significance.

Example 11.2.1 Method Choices

Question: Which of the following is correct and most complete?

- FF is sometimes used to give screening information and for final system choices.
- RSM helps in understanding interaction effects and predicting performance.
- T-testing can, if applied with randomized experimentation, generate strong proof.
- All of the above are correct.
- All are correct except (b) and (d).

Answer: Yes, fractional factorials (FF) are often the last and only design of experiments method used in many projects. Also, modeling the combined effects of factors or “interactions” is possible using response surface methods (RSM). Also, t-testing using randomization can generate proof. Therefore, the correct and most complete answer is (d).

Table 11.1. Brief summary of methods described in this chapter

Method	Advantage	Disadvantage
Two-sample t-tests	Provide a relatively high level of evidence that a single level of a single factor causes a higher average response	Methods only address one factor-at-a-time (OFAT). Compared with screening using fractional factorials, for comparable total costs the Type I and Type II errors are more likely.
Screening using Fractional Factorials (FF)	Provides an inexpensive way to determine which factors from a long list significantly affect system performance. Sometimes, users apply results to support final engineering design decisions	Compared with Response Surface Methods, the methods generate a relatively inaccurate prediction model. Compared with two-sample t-tests, the level of evidence associated with significance claims is subjectively lower.
One-shot Response Surface Methods (RSM)	Create a relatively accurate prediction model and significance information, permitting identifying of interaction effects	Compared with factor screening methods, these methods require substantially larger numbers of experimental runs for a given number of factors.
Sequential Response Surface Methods (RSM)	Generate a relatively accurate prediction model and may require fewer runs than one shot response surface methods.	The derived prediction model will, in general, be less accurate than the one from one-shot response surface methods if the method terminates without using all the runs.
Robust Design based on Profit Maximization (RDPM)	Builds on RSM to directly maximize the sigma level in a cost-effective manner addressing production noise	Complicated; may require substantial experimental cost
Analysis of Variance (ANOVA) followed by multiple t-tests	Offers a standard approach for analyzing significance of factors and/or model terms that addresses the multiplicity of the tests	Compared with Lenth's method and normal probability plots, the Type II errors are generally higher. This is only an analysis method that does not explain which data to collect.

11.3 The Two-sample T-test Methodology and the Word “Proven”

The following class of methods is called “two-sample *t*-testing assuming unequal variances” that can be viewed as the simplest design of experiments methods. Members of this class are distinguished by the initial sample size parameters n_1 and n_2 in *Step 1* and the α level used in *Step 3*.

Roughly speaking, this method is useful for situations in which one is interested in “proving” with a “high level of evidence” that one alternative is better in terms of average response than another. Therefore, there is one factor of interest at two levels. The screening procedure described subsequently can permit several factors to be “proven” significant simultaneously with a comparable number of total tests. However, a subjectively greater level of assumption-making is needed for those screening methods such that the two-sample t-test offers a higher level of evidence.

Definition: The phrase “**blocking factor**” refers to system input variables that are not of primary interest. For example, in a drug study, the names of the people receiving the drug and the placebo are not of primary interest even though their safety is critical.

Algorithm 11.1. Two-sample t-tests

Step 1. a. Develop an experimental table or “DOE array” that describes the levels of all blocking factors and the factor of interest for each run. The ordering of the factor levels should exhibit no pattern, *i.e.*, an effort should be made to allocate all blocking factor levels in an unpatterned way. Ideally, experimentation is “**blind**” so that human participants do not know which level they are testing. Unpatterned ordering can be accomplished by putting n_1 As and n_2 Bs in 1 column on a spreadsheet and pseudo-random uniform [0,1] numbers in the next column. Sorting, we have a “uniformly random” ordering, *e.g.*, 2-1-1-2-2-2-1...

b. Collect $n_1 + n_2$ data, where n_1 of these data are run with factor A at level 1 and n_2 are run with factor A at level 2 following the experimental table.

Step 2. Defining \bar{y}_1 as the average of the run responses with factor A at level 1 and s_1^2 as the sample variance of these responses, and making similar definitions for level 2, one then calculates the quantities t_0 and degrees of freedom (df) using

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad df = \text{round} \left[\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \right] \quad (11.1)$$

where “round” means round the number in brackets to the nearest integer.

Step 3. Find $t_{critical}$ using the Excel formula “=TINV(2*0.05,df)” or using the critical value from a t -table referenced by $t_{\alpha,df}$ (see Table 11.2). If $t_0 > t_{critical}$, then claim that “it has been proven that level 1 of factor A results in a significantly *higher* average or expected value of the response than level 2 of factor A with alpha equal to 0.05”.

Step 4. (Optional) Construct two “box plots” of the response data at each of the two level settings (see below). Often, these plots aid in building engineering intuition.

Table 11.2. Values of $t_{critical} = t_{\alpha,df}$

df	α		
	0.01	0.05	0.1
1	31.82	6.31	3.08
2	6.96	2.92	1.89
3	4.54	2.35	1.64
4	3.75	2.13	1.53
5	3.36	2.02	1.48
6	3.14	1.94	1.44

df	α		
	0.01	0.05	0.1
7	3.00	1.89	1.41
8	2.90	1.86	1.40
9	2.82	1.83	1.38
10	2.76	1.81	1.37
20	2.53	1.72	1.33

Definition: The “median” of m numbers is the $[(m + 1)/2]^{\text{th}}$ highest if m is odd. It is the average of the $(m/2)^{\text{th}}$ highest and the $[(m/2) + 1]^{\text{th}}$ highest if m is even.

Algorithm 11.2. Box and whisker plotting

If the number of data is even, then the 25% (Q1) and 75% (Q3) quartiles are the middle values of the two halves of the data. Otherwise, they are the median including the middle in both halves.

Step 1: Draw horizontal lines at the median, Q1, and Q3.

Step 2: Connect with vertical lines the edges of the Q1 and Q3 lines to form a rectangle or “box”.

Step 3: Then, draw a line from the top middle of the rectangle up to the highest data below $Q3 + 1.5(Q3 - Q1)$ and down from the bottom middle of the rectangle to the smallest observation greater than $Q1 - 1.5(Q3 - Q1)$.

Step 4: Any observations above the top of the upper line or below the bottom of the lower line are called “outliers” and labeled with “*” symbols.

Note that, with only 3 data points, software generally does not follow the above exactly. Instead, the ends of the boxes are often the top and bottom observations.

If we were trying to prove that level 1 results in a significantly *lower* average response than level 2, in *Step 3* of Algorithm 11.1, we would test $-t_0 > t_{critical}$. In general, if the sign of t_0 does not make sense in terms of what we are trying to prove, the above “**one-sided**” testing approach fails to find significance. The phrase “**1-tailed test**” is a synonym for one-sided.

To prove there is any difference, either positive or negative, use $\alpha/2$ instead of α and the test becomes “**two-sided**” or “**2-tailed**”. A test is called “**double blind**” if it is blind and the people in contact with the human testers also do not know which level is being given to which participant. The effort to become double blind generally increases the subjectively assessed level of evidence. Achieving blindness can require substantial creativity and expense.

The phrase “**Hawthorne effect**” refers to a change in average output values caused by the simple act of studying the system, *e.g.*, if people work harder because they are being watched. To address issues associated with Hawthorne

effects and generate a high level of evidence, it can be necessary to include the current system settings as one level in the application of a t-test. The phrase “**control group**” refers to any people in a study who receive the current level settings and are used to generate response data.

Definition: If something is proven using any given α , it is also proven with all higher levels of α . The “**p-value**” in any hypothesis test is the value of α such that the test statistic, *e.g.*, t_0 , equals the critical value, *e.g.*, $t_{\alpha,df}$. The phrase “**significance level**” is a synonym for p-value. For example, if the p-value is 0.05, the result is proven with “alpha” equal to 0.05 and the significance level is 0.05. Generally speaking, people trying to prove hypotheses with limited amounts of data are hoping for small p-values.

Using t-testing is one of the ways of achieving evidence such that many people trained in statistics will recognize a claim that you make as having been “proven” with “objective evidence”. Note that if t_0 is not greater than $t_{critical}$, then the standard declaration is that “significance has not been established”. Then, presumably either the true average of level 1 is not higher than the true average of level 2 or, alternatively, additional data is needed to establish significance.

The phrase “**null hypothesis**” refers to the belief that the factors being studied have no effects, *e.g.*, on the mean response value. Two-sample t-testing is not associated with any clear claims about the factors not found to be significant, *e.g.*, these factors are not proven to be “insignificant” under any widely used conventional assumptions. Therefore, failing to find significance can be viewed as accepting the null hypothesis, but it is not associated with proof.

In general, the testing procedures cannot be used to prove that the null hypothesis is true. The Bayesian analysis can provide “posterior probabilities” or chances that factors are associated with negligible average changes in responses after *Step 1* is performed. This nonstandard Bayesian analysis strategy can be used to provide evidence of factors being unimportant.

11.4 T-test Examples

This section contains two examples, one of which relates to a straightforward application of the t-test method. The second involves answering specific questions based on the concepts. In the first example, an auto company is interested in extending the number of auto bodies that an arc-welding robot can weld without adjustment using a new controller program. The first example is based on the commonly chosen sample size, $n_1 = n_2 = 3$, and selection $\alpha = 0.05$.

If one fails to find significance, that does not mean that the true average difference in responses between the two levels is exactly zero or negative. With additional testing, the test can be re-run and significance might be found. Note that the procedure, if applied multiple times, gives a probability of falsely finding significance (Type I errors) greater than α .

Still, it is common to neglect this difference and still quote the α used in Step 3 as the probability of Type I errors. Therefore, the choice of initial sample size is not critical unless it is wastefully large since additional runs can be added. A

rigorous sequential approach would be to pre-plan on performing at most q sets of runs, with tests after each set, stopping if significance is found. Then, the α used for each test could be α/q such that the overall procedure rigorously guarantees an error rate less than α (e.g., 0.05) using the “Bonferroni inequality” which regulates overall errors.

Table 11.3. One approach to randomize the run order using pseudo-random numbers

Levels	Pseudo-random Uniform Nos.	Run	Level	Sorted Nos.	Response
1	0.583941	1	1	0.210974	$Y_{1,1}=25$
1	0.920469	2	2	0.448561	$Y_{2,1}=20$
1	0.210974	3	1	0.583941	$Y_{1,2}=35$
2	0.448561	4	2	0.589953	$Y_{2,2}=23$
2	0.692587	5	2	0.692587	$Y_{2,3}=21$
2	0.589953	6	1	0.920469	$Y_{1,3}=34$

Algorithm 11.3. First t-test example

- Step 1.* The engineer uses Table 11.3 to determine the run ordering. Pseudo-random uniform numbers were generated and then used to sort the levels for each run. Then, we first input level 1 (the new additive) into the system and observed the response 25. Then, we input level 2 (the current additive) and observed 20 and so on.
- Step 2.* Responses from welding tests are shown in the right-hand column of Table 11.3. The engineer calculated $\bar{y}_1 = 31.3$, $\bar{y}_2 = 21.3$, $s_1^2 = 30.3$, $s_2^2 = 2.33$, $t_0 = 3.03$, and $df = 2$.
- Step 3.* The critical value given by Excel “=TINV(0.1,2)” was $t_{critical} = 2.92$. Since t_0 was greater than $t_{critical}$, we declared, “We have proven that level 1 results in a significantly higher average mean value than level 2 with alpha equal to 0.05.” The p-value is 0.047.
- Step 4.* A box plot from Minitab® software is below which shows that level 1 results in higher number of bodies welded on average. Note that with 3 data Minitab® defines the lowest data point at Q1 and the highest data point as Q3.

Example 11.4.1 Second T-test Application

A work colleague wants to “prove” that his or her software results in shorter times to register the product over the internet on average than the current software. Suppose six people are available for the study: Fred, Suzanne,...(see below).

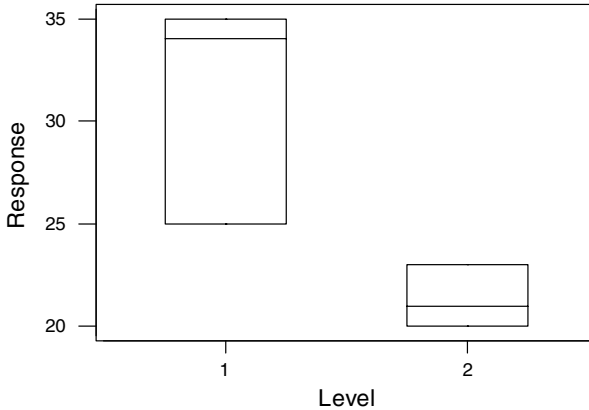


Figure 11.1. Minitab[®] box plot and whisker for the autobody welding example

Question 1: How many factors, response variables, and levels are involved?

Answer 1: There are two correct answers: (1) two factor (software) at two levels (new and old) and 1 response (time) and (2) two factors (software and people) at two and six levels and 1 response (time). If the same person tested more than one software, people would be a factor.

Question 2: What **specific** instructions (that a technician can understand) can you give her to maximize the level of evidence that she can obtain?

Answer 2: Assume that we only want one person to test one software. Then, we need to randomly assign people to levels of the factor. Take the names Fred, Suzanne,... and match each to a pseudo-random number, *e.g.*, Fred with 0.82, Suzanne with 0.22,... Sort the names by the numbers and assign the top half to the old and the bottom half to the new software. Then, repeat the process with a new set of pseudo-random numbers to determine the run order. There are other acceptable approaches, but both assignment to groups and run order must be randomized.

Question 3: In this question, the following data is needed:

New software	Old software
Fred – 35.6 sec	Juliet – 45.2 sec
Suzanne – 38.2 sec	Bob – 43.1 sec
Jane – 29.1 sec	Mary – 42.7 sec

Analyze the above data and draw conclusions that you think are appropriate.

Answer 3: We begin by calculating the following: $\bar{y}_1 = 34.30$, $\bar{y}_2 = 43.67$, $s_1^2 = 21.97$, $s_2^2 = 1.80$, $t_0 = 3.33$, and $df = \text{round}[2.3] = 2$. Note that we are hoping the

average time is lower (a better result), therefore the sign of t-critical makes sense and we can ignore it for the calculation. Since $3.33 > 2.92$ we have proven that the new software reduces the average registration time with $\alpha = 0.05$.

Question 4: How might the answer to the previous question support decision-making?

Answer 4: The software significantly reduces average times, but that might not mean that the new software should be recommended. There might be other criteria such as reliability and cost of importance.

11.5 Randomization and Evidence

One activity is common to all of the applications of the design of experiments (DOE) methods in this book. This activity is “**randomization**” which is the allocation of blocking factor levels to runs in a random or unpatterned way in experimental planning. For example, the run order can be considered to be a blocking factor. The act of scrambling the run order is a common example of randomization. Also, the assignment of people and places to factor levels can be randomized.

Philosophically, the application of randomization is critical for proving that certain factor changes affect average response values of interest. Many experts would say that empirical proof is impossible without randomization. Data collection is called an “experiment” if randomization is used and an “observational study” if it is not. Further, many would say experiments are needed for “doing science” although science is also associated with physics-based modeling.

Note that attempts to control usually uncontrollable factors during experimentation can actually work against development of proof, because control can change the system so that proof derived (if any) pertains to a system that is different than the one of interest. Often, the process is aided through the creation of an experimental plan or table showing the levels of the factor and the blocking factors (if any). The use of a planning table is illustrated (poorly) in the next example.

Example 11.5.1 Poor Randomization and Waste

Question 1: Assume that the experimental designer and all testers are watching all trials related to Table 11.4. The goal of the new software is task time reduction. Which is correct and most complete?

- The data can be used to prove the new software helps with $\alpha = 0.05$.
- The theory that the people taking the test learned from watching others is roughly equally plausible to the theory that the new software helps.
- The theory that women are simply better at the tasks than men is roughly equally plausible to the theory that the new software helps.
- The tests would have been much more valuable if randomization had been used.

- e. All of the above are correct except (a).

Answer 1: The experimental plan has multiple problems. The run order is not randomized so learning effects could be causing the observed variation. The assignment of people to levels is not randomized so that gender issues might be causing the variation. The test was run in an unblind fashion, so knowledge of the participants could bias the results. Therefore, the correct answer is (e).

Table 11.4. Hypothetical example in which randomization is not used

Run (blocking factor)	Software	Tester (blocking factor)	Average time per task
1	Old	Jim	45.2
2	Old	Harry	38.1
3	Old	George	32.4
4	New	Sue	22.1
5	New	Sally	12.5
6	New	Mary	18.9

Question 2: Which is correct and most complete?

- Except for randomization issues, t-testing analysis could be reasonably used.
- $t_0 = 4.45$ for the two sample analysis of software, assuming unequal variances.
- The experiment would be “blind” if the testers did not know which software they were using and could not watch the other trials.
- All of the above are correct.
- None of the above is correct.

Answer 2: Often, in experimentation using t-testing, there are blocking factors that should be considered in planning and yet the t-testing analysis is appropriate. Also, the definition of blind is expressed in part (c). Therefore, the answer is (d).

11.6 Errors from DOE Procedures

Investing in experimentation of any type is intrinsically risky. This follows because if the results were known in advance, experimentation would be unnecessary. Even through competent application of the methods in this book, errors of various types will occur. Probability theory can be used to predict the chances and/or magnitudes of different errors as described in Chapter 19. The theory can also aid in the comparison of method alternatives.

In this section, concepts associated with errors in testing hypotheses are described which are relevant to many design of experiments methods. These

concepts are helpful for competent application and interpretation of results. They are also helpful for appreciating the benefits associated with standard screening using fractional factorials.

Table 11.5 defines Type I and Type II errors. The definition of these errors involves the concepts of a “true” difference and absence of the true difference in the natural system being studied. Typically, this difference relates to alternative averages in response values corresponding to alternative levels of an input factor. In real situations, the truth is generally unknown. Therefore, Type I and Type II errors are defined in relation to a theoretical construct. In each hypothesis test, the test either results in a declaration of significance or failure to find significance.

Table 11.5. Definitions of Type I and Type II errors

		Nature or truth	
		No difference exists	Difference exists
Declaration	Significance is found	Type I error	Success
	Failure to find	Semi-success	Type II error

Failure to find significance when no difference exists is only a “semi-success” because the declaration is indefinite. Implied in the declaration is that with more runs or slightly different levels, a difference might be found. Therefore, the declaration in the case of no true difference is not as desirable as it could be.

As noted previously, theory can provide some indication of how likely Type I and Type II errors are in different situations. Intuitively, for two-sample t-testing, the chance of errors depend on all of the following:

- The sample sizes used, n_1 and n_2
- The α used in the analysis of results
- The magnitude of the actual difference in the system (if any)
- The sizes of the random errors that influence the test data (caused by uncontrolled factors)

Like many testing procedures, the two-sample t-test method is designed to have the following property. For testing with chosen parameter α and any sample sizes, the chance of Type I error equals α . In one popular “frequentist” philosophy, this can be interpreted in the following way. If a large number of applications occurred, Type I errors would occur in α fraction of these cases. However, the chance of a Type I equaling α is only precisely accurate for specific assumptions about the random errors described in Chapter 19.

Therefore, fixing α determines the chance of Type I errors. At the same time, the chance of a Type II error can, in general, be reduced through increasing the sample sizes. Also, the larger the difference of interest, the smaller the chance of Type II error. In other words, if the tester is looking for large differences only, the chance of missing these distances and making a Type II error is small, in general.

Example 11.6.1 Testing a New Drug

Question: An inventor is interested in testing a new blood pressure medication that she believes will decrease average diastolic pressure by 5 mm Hg. She is required by the FDA to use $\alpha = 0.05$. What advice can you give her?

- Use a smaller α ; the FDA will accept it, and the Type II error chance is lower.
- Budgeting for the maximum possible sample size will likely help prove her case.
- She has a larger chance of finding a smaller difference.
- All of the above are correct.
- All are correct except (b) and (d).

Answer: As noted previously, if something is proven using any given α , it is also proven with all higher levels of α . Therefore, the FDA would accept proof with a lower level of α . However, generally proving something for a lower α implies an increased chance of Type II error. Generally, the more data, the more chance of proving something that is true. Also, finding smaller differences is generally less likely. Therefore, the correct answer is (b).

11.7 Chapter Summary

This chapter has provided an overview of the design of experiments (DOE) methods in this book. To simplify, fractional factorial methods are useful for screening to find which of many factors matter. Response surface methods (RSM) are useful for developing relatively accurate surface predictions, including predicting so-called interactions or combined effects of factors on average responses. Sequential RSM offer a potential advantage in economy in that possibly fewer runs will be used. Robust design methods address the variation of uncontrollable factors and deliver relatively trustworthy system design recommendations.

The method of t-testing was presented with intent to clarify what randomization is and why it matters. Also, t-testing was used to illustrate the use of information to support method related decision-making, *e.g.*, about how many test runs to do at each level. Theoretical information is presented to clarify the chances of different types of errors as a function of method design choices.

Finally, the concept of randomization is described, which is relevant to all of the appropriate application of all of the design of the experiments methods in this book. Randomization involves a careful step of planning experimentation that is critical for achieving proof and high levels of evidence.

The following example illustrates how specific key input variables and DOE methods can be related to real world problems.

Example 11.7.1 Student Retention Study

Question 1: Suppose that you are given a \$4,000,000 grant over three years to study retention of students in engineering colleges by the Ohio State Board of Regents. The goal is to provide proven methods that can help increase retention, *i.e.*, cause a higher fraction of students who start as freshmen to graduate in engineering. Describe one possible parameterization of your engineered system including the names of relevant factors and responses.

Answer 1: The system boundaries only include the parts of the Ohio public university network that the Board of Regents could directly influence. These regents can control factors including: (1) the teaching load per faculty (3 to 7 course per year), (2) the incentives for faculty promotion (weighted towards teaching or research), (3) the class size (relatively small or large), (4) the curriculum taught (standard or hands-on), (5) the level of student services (current or supplemented), and (6) the incentives to honors students at each of the public campuses (current or augmented). Responses of interest include total revenues per college, retention rates of students at various levels, and student satisfaction ratings as expressed through surveys.

Question 2: With regard to the student retention example, how might you spend the money to develop the proof you need?

Answer 2: Changing university policies is expensive. Expenses would be incurred through additions to students' services for selected groups, summer salary for faculty to participate, and additional awards to honors students. Because of the costs, benchmarking and regression analyses techniques applied, using easily obtainable data would be relevant. Still, without randomized experimentation, proof of cause and effect relationships relevant to Ohio realities may be regarded as impossible. Therefore, I would use the bulk of the money to perturb the existing policies. I would begin by dividing the freshman students in colleges across the state into units of approximately the same size in such a way that different units would naturally have minimal interaction. Then, I would assign combinations of the above factor levels to the student groups using random numbers to apply standard screening using fractional factorials with twelve runs ($n = 12$). I would evaluate the responses each year associated with the affected student groups applying the fractional factorial analysis method. As soon as effects appeared significant, I would initiate two-sample t-tests of the recommended settings vs the current using additional groups of students to confirm the results, assuming the remaining budget permits. The fractional factorial and added confirmation runs would likely consume several million dollars. However, it seems likely that the findings would pay for themselves, because the state losses per year associated with poor retention have been estimated in the tens of millions of dollars. These costs do not include additional losses associated with university ratings stemming from poor retention.

11.8 Problems

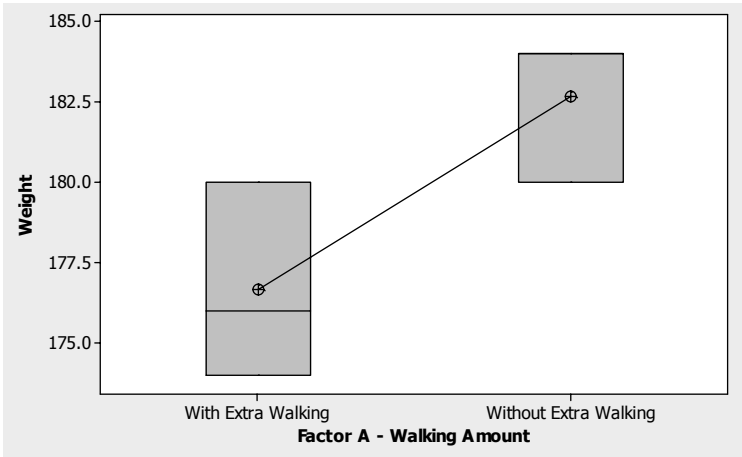
1. Consider applying DOE to improving your personal health. Which of the following is correct and most complete?
 - a. Input factors might include weight, blood pressure, and happiness score.
 - b. Output responses might include weight, blood pressure, and happiness score.
 - c. Randomly selecting daily walking amount each week could generate proof.
 - d. Walking two months 30 minutes daily followed by two months off can yield proof.
 - e. Answers to parts (a) and (d) are both correct.
 - f. Answers to parts (b) and (c) are both correct.

2. Which is a benefit of DOE in helping to add definitiveness in design decision-making?
 - a. Engineers feel more motivated because proof is not needed for changes.
 - b. Tooling costs are reduced since dies must be designed and built only once.
 - c. Carefully planned input patterns do not support authoritative proof.
 - d. Documentation becomes more difficult since there is no moving target.
 - e. Quality likely improves because randomization does not generate rigorous proof.

3. Based on Chapter 1, which of the following is correct and most complete?
 - a. Taguchi contributed to robust design and Box co-invented FF and RSM.
 - b. Deming invented FF and invented RSM.
 - c. Shewhart invented ANOVA.
 - d. All of the above are correct.
 - e. All of the above are correct except (a) and (d).

4. Which of the following is correct and most complete?
 - a. FF is helpful for finding which factors matter from a long list with little cost.
 - b. RSM helps in fine tuning a small number of factor settings.
 - c. Robust engineering helps in that it is relatively likely to generate trustworthy settings.
 - d. All of the above are relevant advantages.
 - e. All of the above are correct except (a) and (d).

Data from Figure 11.2 will be used for Questions 5 and 6.



Run	Weight	Factor A – Walking amount
1	184	Without extra walking
2	180	With extra walking
3	184	Without extra walking
4	176	With extra walking
5	174	With extra walking
6	180	Without extra walking

Figure 11.2. Data and Minitab® Box and Whisker plot for weight loss example

5. Which of the following is correct and most complete?
 - a. $t_0 = 3.7$, which is “>” the relevant critical value, but nothing is proven.
 - b. $t_0 = 2.7$, we fail to find significance with $\alpha = 0.05$, and the plot offers nothing.
 - c. $t_0 = 2.7$, which is significant with $\alpha = 0.05$, we can claim proof walking helps.
 - d. The run ordering is not random enough for establishing proof.
 - e. All of the above are correct except (a).

6. Calculate the degrees of freedom (df) using data from the above example.

7. Consider the Second Two-sample t-test example in Section 11.4.1 of this chapter. Assume that no more tests were possible. Which is correct and most complete?
- Randomly assigning people to treatments *and* run order is essential for proof.
 - It is likely true that failing to find significance would have been undesirable.
 - The test statistic indicates a real difference larger than the noise is present.
 - Significance would also have been found for any value of α larger than 0.05.
 - Answers to parts (a) and (d) are both correct.
 - All of the above answers are correct.
8. Assume you t-test with $n_1 = n_2 = 4$. Which is correct and most complete?
- Using $n_1 = n_2 = 3$ would likely reduce the chance of Type I and Type II errors.
 - The chance of finding significance can be estimated using theory.
 - Random assignment of run ordering makes error rate (both Type I and Type II probabilities) estimates less believable.
 - Finding significance guarantees that there is a true average response difference.
 - All answers except (d) are correct.

Use the following design of experiments array and data to answer Questions 9 and 10. Consider the following in relation to proving that the new software reduces task times.

Table 11.6. Software testing data

Run	Software	Tester	Average time per task
1	Old	Mary	45.2
2	New	Harry	38.1
3	Old	George	32.4
4	New	Sue	22.1
5	New	Sally	12.5
6	Old	Phillip	18.9

9. Which is correct and most complete?
- The above is an application of a within-subjects design.
 - The above is an application of a between-subjects design.
 - One fails to find significance with $\alpha = 0.05$.
 - The degrees of freedom are greater than or equal to three.

- e. All of the above are correct.
 - f. All of the above are correct except (a) and (e).
10. Which is correct and most complete?
- a. The blocking factor tester has been randomized over.
 - b. $t_0 = 1.45$ for the two sample analysis of software assuming unequal variances.
 - c. Harry, Sue, and Sally constitute the control group.
 - d. All of the above are correct.
 - e. All of the above are correct except (a) and (e).
11. Which is correct and most complete for t-testing or factor screening?
- a. In general, adding more runs to the plan increases many types of error rates.
 - b. Type I errors in t-testing include the possibility of missing important factors.
 - c. Type II errors in t-testing focus on the possibility of missing important factors.
 - d. Standard t-testing can be used to prove the insignificance of factors.
 - e. All of the above are correct.
 - f. All of the above are correct except (a) and (e).