Why the Results of Designed Experiments are Often Disappointing: A Case Study

J. Clifton Young
Univeristy of Waterloo

RR-94-10 (NT)
October 1994

Why the Results of Designed Experiments Are Often Disappointing: A Case Study Illustrating the Importance of Blocking, Replication and Randomization

J. Clifton Young
Institute for Improvement in Quality and Productivity
University of Waterloo

Abstract

A major cause of disappointing results and a corresponding decline in interest in the use of designed experiments is the lack of expert initial planning. In addition to the careful choice of factors and levels, three essential components of a well planned (designed) experiment are the techniques of blocking, replication and randomization. The impact of these techniques in improving both the effectiveness and efficiency of planned experiments is illustrated using a case study on the closing effort of the hatch of a car.

Key Words: Factorial Experiments, Fractional Factorial Experiments, Blocking, Randomization, Replication, Experimental Design, Data Collection.

Introduction

Having been involved in the teaching and application of statistical experimental design for over 25 years, I was very pleased, during the '80's, to see the growth in interest in the application of this highly efficient methodology to the solution of industrial problems. I am now very concerned to see a corresponding decline in interest over the past few years.

There are several reasons for this waning enthusiasm. Firstly, as with the earlier enthusiasm for control charts, a lot of people jumped on the "DOE" bandwagon, naively thinking (or having been told) that this "new" technique would solve all their problems. Needless to say it did not. In fact, strict adherence to standard operating procedure and straightforward analysis based on the stratification of data by shift, machine, palette, raw material lot, filler head, cavity and so on will solve a large number of the problems that plague industry.

Even when there is the potential for the solution of a problem by use of a well designed experiment, the method has often failed or lead people in the wrong direction. Why is this? One reason is that people undertake to do a "DOE" or a "Design of Experiment" without thinking about the actual meaning of the words they are using, especially the critical word "design". This word reflects the importance of careful up-front planning in the experimental process; too often this critical stage is given far too little attention. For example, critical decisions on what factors should be included in the experiment, what levels should be used and whether the levels of one factor (say line speed in a continuous heat soak operation) should depend on another (say temperature level) are often given far too little attention. These decisions can only be made by a carefully selected team with experience in operations, engineering, and statistics. Further discussion of these aspects of experimental planning can be found in Young (1989).

This case study illustrates the use of blocking, replication and randomization, three critical components of well-designed experiments that are often missing or glossed over in the usual industrial short courses on experimental design. In fact, apart from some rather mechanical instruction on the use of orthogonal arrays or their equivalent, many such courses are mainly concerned with experimental analysis rather than design. This is very unfortunate since, if factors and levels are well chosen, careful blocking, replication and randomization will almost always greatly improve the precision of any experiment as well as guarantee the validity or accuracy of the corresponding conclusions. These techniques will be illustrated in the following case study from a car assembly plant.

The Problem

During the initial production of the hatch-back version of a new design of a car, the effort needed to close the hatch was unstable; although the required closing effort was often perfectly acceptable, it was often excessive.

Although there were many possible explanations for the problem, there was a strong desire to investigate the influence on process stability of a component of the hatch gasket called the stuffer. In particular there was concern that variation relative to nominal length, firmness and location might cause problems. Since there were no scientifically determined specification limits for these factors, levels were chosen at the high and low end of the observed variation. Thus, in addition to determining the influence on closing effort of these three factors, another goal of the experiment was to determine acceptable specifications.

Experimental Planning Considerations: The Need for Precision and Accuracy

In order to keep the required number of observations down, a two-level factorial experiment was chosen. This entailed running all eight combinations of the three factors as indicated in Table 1. For each factor, the negative sign represents a stuffer at the low end of normal variation in placement, thickness or firmness and the plus sign the high end.

Table 1: The basic design.
Note that the traditional method of "+" and "-" signs has been used to denote the high and low levels of the factors. The numbers 1 and 2 could have been used just as well.

	Factor		
Run	**Length (L)**	**Position (P)**	**Firmness (F)**
1	-	-	-
2	+	-	-
3	-	+	-
4	+	+	-
5	-	-	+
6	+	-	+
7	-	+	+
8	+	+	+

The eight desired treatment combinations were obtained by sorting stuffers as they came in from the supplier.

Although this design is very efficient in that it provides estimates of the three main effects and the three two-factor interactions, in its unreplicated form there is no effective measure of error. The three-factor interaction can usually be safely assumed to be negligible and used but, since it provides only one degree of freedom of information on error, it is not very sensitive to detecting anything but very large effects. The alternative of using a probability plot (Box, Hunter and Hunter, 1978, Section 10.9) to determine the significant effects is unreliable in view of the considerable variation found in such plots for only seven points (Daniel and Woods, 1980, Appendix 3a). Even if there were an effective error term, small experiments will usually result in poor precision.

In order to be significant at approximately the five percent level, an effect must be roughly two standard errors in magnitude. Thus the standard error (SE) is an excellent indicator of the precision of an experiment (Cox, 1958, Section 8.2). The general formula for the SE for comparing two means is

$$SE\,(\bar{y}_1 - \bar{y}_2) = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\,, \qquad [1]$$

where

s is the experimental error. In a simple experiment this can be approximated by the overall process standard deviation.

n_i is the number of independent runs in the experiment at the i^{th} level of a given factor (the number of observations in $\bar{y}_i$).

$\bar{y}_i$ is the mean of all n_i observations at the i^{th} level of the factor.

In an eight-run experiment with all factors at two levels (n_1=n_2=4)

$$SE\,(\bar{y}_{High} - \bar{y}_{Low}) = \frac{s}{\sqrt{2}}$$

If the experiment is spread over a period of time that is representative of normal operating conditions, and this is absolutely necessary if the conclusions are to be accurate, s will usually be large and the resulting precision poor. In fact, a large s is often the primary reason for experimenting.

In planning the hatch experiment, the original idea was to make each of the eight runs on a different car and then replicate until the desired precision was achieved. The model for such an experiment is

$$y_{ijkq} = m + l_i + p_j + (lp)_{ij} + f_k + (lf)_{ik} + (pf)_{jk} + (lpf)_{ijk} + \theta_{ijkq} \qquad [2]$$

where

y_{ijkq} is the observed force on the car $ijkq$. ($i, j, k = 1, 2$ and $q = 1,...Q,$ the number of runs (replicates) at each of the eight experimental conditions).

m is the overall mean.

l_i, p_j and f_k are the length, position and firmness main effects respectively.

$(lp)_{ij}$, $(lf)_{ik}$, etc. are the corresponding interaction effects.

θ_{ijkq} is the experimental error term with distribution $N(0, \sigma^2)$.

σ is the car-to-car standard deviation for the process.

Knowledge of the overall process standard deviation can be used as a rough estimate of σ. This can then be inserted in the formula for the SE to determine the size of experiment that will result in the required precision. In this case, since the engineers wanted to be fairly certain of estimating effects to within approximately plus or minus one unit, they decided to try for a SE of 0.5 (to ensure that an estimated effect of 1 unit would be significant at about the 5% level). Substituting this value in equation 1, using the current process value of $s = 4.61$ for car-to-car variation, and solving for $n = n_1 = n_2$ results in $n = 170$. Clearly this is far too large.

It must be noted that it is not valid to increase n_i by taking several repeat observations during a single run of an experiment. An eight-run experiment can provide only eight independent observations with the result that $n_i = 4$ for comparing means. For instance, setting up a car with a stuffer with a specific firmness, length and location and then taking five observations of closing effort provides one independent observation (the mean of the five observations) not five. Similarly, several observations on a single part or several parts in a row during a single run of a process provides only one independent observation for analysis purposes. It may, of course, be of interest to include the variation among such observations as another experimental response variable. Further discussion of independence can be found in Cox (1958, Chapter 1).

A much better design is to block the experiment such that a whole set of eight runs is carried out on one car. This results in very high precision since car-to-car variation now enters as a term (b_q) in the statistical model as follows:

$$y_{ijkq} = m + l_i + p_j + (lp)_{ij} + f_k + (lf)_{ik} + (pf)_{jk} + (lpf)_{ijk} + b_q + \epsilon_{ijkq} \qquad [3]$$

where

b_q is the block (car) effect ($q = 1, \ldots Q,$ the number of cars).

ϵ_{ijkq} is the new error term ($\epsilon_{ijkq} \sim N(0, \delta^2)$ where δ is the *within* car standard).

Since the b_q term pulls car-to-car variation out of the error term in equation 2, blocking leads to a considerable reduction in experimental error. The within-car variation $s = .879$ currently observed in the process can now be used in equation 1 to determine sample size. Substituting this value and solving for $SE = .5$ results in $n = 6.2$. This means that two sets of eight runs should give the desired level of precision.

The drawback is that two cars cannot be regarded as particularly representative of production. This drawback was overcome to some extent by using the signs ("levels") of the three factor interaction *LPF* to set up blocks of size four (see Table 2). Thus all runs with "-" signs in the *LPF* interaction were run on one day by one operator on the first car, those with "+" signs by another operator on a second day on car two, with cars three and four used in a similar manner for the last two blocks. Note that all runs on a given car (block) were carried out in a single day by one operator. This was done because the more non-experimental factors held constant within a block the higher the resulting precision.

The model for the revised experiment remains essentially the same as shown in [3] except that the $(lpf)_{ijk}$ term is dropped and $q=1, 2, 3, 4$ corresponding to the car the observation was collected from. Note that not all combinations *i, j, k, q* are included in this design.

The benefits of blocking are substantial:

1. Blocking results in the desired precision with far fewer observations (16 instead of 340).

2. The relatively small blocks used enable the inclusion of four cars in the experiment. This provides replication (representivity) and hence improves the accuracy or reliability of the conclusions.

3. Since only four runs had to be performed per day, the experiment was much less intrusive on the operator's time. He thus had time to carry out his normal duties with relatively little pressure. In other situations, the relatively low number of runs per day means more time to reach steady state and more time to ensure that the experimental runs are carried out exactly as intended. This is especially important with hard-to-change variables such as temperatures and viscosities.

4. If there is a problem with the execution of the runs in any block, they can be run over again using another car on another day. Since any systematic difference between the two days is removed by the block effect, this will not introduce any bias into the experiment.

5. If, after the analysis of the experiment is carried out, the results look promising but not definitive, two more blocks of four can be added in order to increase precision (to $SE\,(\bar{y}_{high} - \bar{y}_{low}) = s/\sqrt{6}$).

6. Since most processes are relatively stable in the short term (within a block), a blocked experiment not only results in relatively high precision for estimating the effects of the factors of interest, but also in an estimate of the degree of instability due to non-experimental factors (those factors that vary from block to block).

7. Since four separate runs can be carried out on a car fairly quickly, final shipment to the customer of that car was never held up more than part of a day.

The final component of the design is the randomization of the order of the four runs in each block. This provides insurance against systematic bias creeping in and leading to incorrect conclusions.

Table 2: The experimental layout in blocks with the resulting hatch closing efforts and the randomized run order in brackets.

Run Number	Effects							Closing Effort (CE)		Treatment
(Table 1)	L	P	LP	F	LF	PF	LPF	Experimental Results		Means
								Block 1	Block 3	
1	-	-	+	-	+	+	-	7.0 (2)	7.0 (4)	7.0
4	+	+	+	-	-	-	-	7.0 (4)	7.5 (2)	7.25
6	+	-	-	+	+	-	-	6.5 (1)	7.0 (3)	6.75
7	-	+	-	+	-	+	-	7.0 (3)	4.5 (1)	5.75
								Block 2	Block 4	
2	+	-	-	-	-	+	+	5.0 (1)	14.0 (4)	9.5
3	-	+	-	-	+	-	+	5.0 (3)	11.0 (2)	8.0
5	-	-	+	+	-	-	+	3.5 (2)	13.0 (1)	8.25
8	+	+	+	+	+	+	+	4.5 (4)	13.5 (3)	9.0
est. effects*	.875	-.375	.375	-.500	0	.250	**			

* Average difference between the means at the high and low levels of each factor
** Not meaningful since it is the difference between the averages of cars 1 & 3 and cars 2 & 4

The final experimental plan is presented in Table 2 with the resulting closing efforts and run order indicated in brackets beside the data. The observed range of data (3.5 to 14) was felt to be quite typical of the process. This is very reassuring as far as the representivity of the experiment is concerned.

In Table 2 the runs listed in Table 1 are rearranged to facilitate identification of the blocks (cars). Note that the three-factor interaction *LPF* corresponds to differences between cars. Since every one of the other six effects has two observations at each of the high and low levels in each block, these six effects are orthogonal to cars (balanced within cars) and will be unaffected by car-to-car variation.

The Analysis

The analysis of variance (ANOVA) for the closing effort (CE) data in Table 2 is given in Table 3. The format is typical of most standard statistical analysis programs.

Table 3: Analysis of variance of the closing effort data from Table 2.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	162.000	18.000	15.57	0.0017
Error	6	6.938	1.156		
Corrected Total	15	168.938			
	R-Square		Root MSE		
	0.959		1.075		

Source	DF	Treatment SS	Mean Square	F Value	Pr > F
Car (Blocks)	3	156.563	52.188	45.14	0.0002
Firmness (F)	1	1.000	1.000	0.87	0.3883
Position (P)	1	0.563	0.563	0.49	0.5116
Length (L)	1	3.063	3.063	2.65	0.1548
Firm*Pos (FP)	1	0.250	0.250	0.22	0.6583
Firm*Len (FL)	1	0.000	0.000	0.00	1.0000
Pos*Len (PL)	1	0.563	0.563	0.49	0.5116

Since the root MSE (root mean squared error) is small (s = *1.075* with 6 d.f.) compared to normal process variation (normally in the range 2 to 19), the statistical model (equation [3]) has explained most of the process variation. (Note that the experimental s is close to the within-car variation (s = *.879*) used to determine sample size). In fact, the experiment explains 95.9% (R^2 = .959) of the experimental variation. Essentially all of this variation is, however, explained by the overwhelmingly significant (p = *.0002*) car-to-car variation. The proportion of variation explained by car-to-car variation is given by

$$\% \ Car\text{-}to\text{-}Car \ Variation = 100 \times \frac{SS(cars)}{Corrected \ SS}$$

$$= 100 \times \frac{156.563}{168.938}$$

$$= 92.7\%$$

This is a reflection of car-to-car instability in the process due to factors other than those that were controlled in the experiment. No other factor was significant.

It sometimes happens that, although not statistically significant, a factor may be regarded as promising in view of its magnitude. The largest effect in this experiment is length with

$$\bar{y}_{long} - \bar{y}_{short} = 8.125 - 7.25 = .875 .$$

Note that this is identical to the *L* effect in Table 2. As discussed earlier,

$$SE\,(\bar{y}_{long} - \bar{y}_{short}) = s\sqrt{\frac{1}{8} + \frac{1}{8}} = \frac{s}{2} = .538 .$$

Thus a 95% confidence interval for the effect of moving from the low end to the high end of specifications is

$$\bar{y}_{long} - \bar{y}_{short} \pm t_{(6)}\, SE\,(\bar{y}_{long} - \bar{y}_{short})$$

or

$$0.875 \pm 2.447 \times .538 = 0.875 \pm 1.317$$

which gives the range

(-.442, 2.194).

Thus, lengths consistently near the high end of specifications may increase the average closing effort by as much as 2.19 units (but might decrease it by up to 0.442 units). If a change in mean of about 2 units is of practical importance, it is worth getting more data to increase precision and thus narrow the width of the confidence interval. How much more data is needed to achieve the desired precision can be found by trying various values of *n* in the standard error formula. In view of the fact that the expected mean change was 0.875 and especially that 92.7% of the experimental variation is due to factors other than those associated with the stuffer, it was decided to look elsewhere.

Reanalysis Ignoring the Blocking

A reasonable guess as to what might have happened if the experiment had not been blocked but had been performed on 16 randomly chosen cars, can be obtained by lumping the *SS(Car)* together with the *SS(Error)* to give a modified

$$\begin{aligned} SS'(Error) &= 156.5625 + 6.9375 \\ &= 163.500 \end{aligned} \qquad \text{(9 df)}$$

The resulting ANOVA table is given in Table 4.

Table 4: ANOVA table for the hypothetical 16 run experiment (blocking ignored)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	5.438	0.906	0.05	0.9992
Error	9	163.500	18.167		
Corrected Total	15	168.938			
	R-Square		Root MSE		
	0.032		4.262		

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Firmness (F)	1	1.000	1.000	0.06	0.8198
Position (P)	1	0.563	0.563	0.03	0.8642
Length (L)	1	3.063	3.063	0.17	0.6910
Firm*Pos (FP)	1	0.250	0.250	0.01	0.9092
Firm*Len (FL)	1	0.000	0.000	0.00	1.0000
Pos*Len (PL)	1	0.563	0.563	0.03	0.8642

Not surprisingly, none of the experimental effects is significant. On a superficial basis this seems to agree with the result in the previous section. There is, however, a major difference: this time the root MSE is $s' = 4.262$ as compared to $s = 1.075$. This compares closely to the overall process $s = 4.61$ discussed earlier. As well, only 3.2% of the variation has been explained. A 95% confidence interval from the unblocked experiment would have width

$$\pm\, t(9)\, s'\sqrt{\frac{1}{8}+\frac{1}{8}} = \pm\, 2.262 \times \frac{4.262}{2}$$

$$= \pm\, 4.816$$

as compared to 1.317. Clearly, the precision is much poorer. A 95% confidence interval for the effect of length would be

$$0.875 \pm 4.816 \quad \text{or} \quad (-3.941,\ 5.691).$$

Although this is certainly not significant, we are now, because of the poor precision, left with the much less satisfying conclusion that the effect could be anywhere in this much wider range.

Blocking in an Unreplicated Experiment

An indication of the fact that blocking can play a useful role even in an unreplicated experiment can be gained by ignoring the results from blocks 1 and 2 and analyzing blocks 3 and 4 as though they comprised the entire experiment. These two blocks have been chosen because they cover all eight combinations of the three factors under study and because the relatively large difference between cars three and four (blocks) best illustrates the benefits of blocking in this situation.

The column labeled "blocked" in the "Experimental Results" column of Table 5 contains the data; the estimated effects are in the row labeled "Blocked" below the contrast coefficients. A probability plot (Box, Hunter & Hunter, Section 10.9) of the estimated effects is given in Figure 1.

Table 5: Contrast Analysis for various possible experimental results from unreplicated eight-run experiments

								Experimental Results				
Run Number	Effects							Blocked	Random			
(Table 1)	L	P	LP	F	LF	PF	LPF		1	2	3	4
1	-	-	+	-	+	+	-	7.0	13.5	7.0	11.0	4.5
4	+	+	+	-	-	-	-	7.5	13.0	7.5	7.0	7.5
6	+	-	-	+	+	-	-	7.0	7.0	14.0	4.5	7.0
7	-	+	-	+	-	+	-	4.0	4.5	11.0	13.5	11.0
2	+	-	-	-	-	+	+	14.0	11.5	4.5	13.0	7.0
3	-	+	-	-	+	-	+	11.0	14.5	13.5	14.0	13.0
5	-	-	+	+	-	-	+	13.0	7.5	13.0	7.0	14.0
8	+	+	+	+	+	+	+	13.5	7.0	7.0	7.5	13.5
	Estimated Effects											
Blocked	1.163	-1.125	1.125	-.375	-.125	.125	6.375					
Random 1	-.375	-.125	1.125	-6.375	1.375	-1.375	.375					
Random 2	-2.875	.125	-2.125	3.125	1.375	-4.625	-.375					
Random 3	-3.375	1.625	-3.125	-3.125	-.875	3.125	1.375					
Random 4	-1.875	3.125	.375	3.375	-.375	-1.375	4.375					

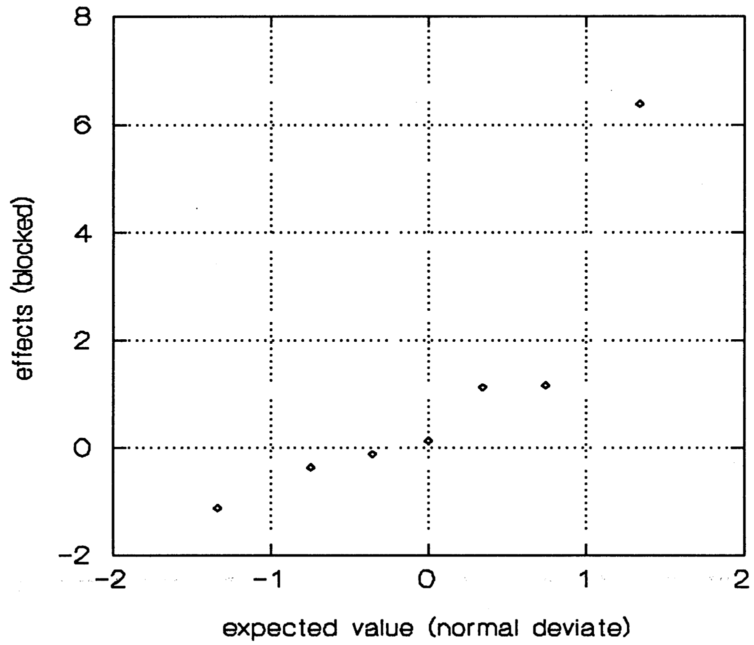


Figure 1: Probability plot of the estimated effects from an unreplicated, blocked design (cars 3 and 4).

The block (*LPF*) effect ($\bar{y}_{car\ 4} - \bar{y}_{car\ 3} = 6.375$) stands out clearly from the rest; it is more than three times as large as any of the other six effects. None of the other effects stand out. In this case, the conclusion would be essentially the same as in the fully replicated experiment discussed earlier. However, the decision clearly depends much more on individual judgement of the appearance of the probability plot than does the much more objective significance test of the earlier section. For this experiment, there would be little disagreement regarding the conclusion. In general, although effective with larger experiments, probability plots from small samples are too erratic to give reliable conclusions with eight- or nine-run experiments. This is well illustrated by the plots in Appendix 3a of Daniel and Woods (1980).

In this case one can use the non-significant effects to estimate *s*

$$s = \sqrt{\frac{2 \times \sum (\textit{non-significant effects})^2}{\textit{number of non-significant effects}}}$$

$$= \sqrt{\frac{2 \times (1.163^2 + (-1.125)^2 + \ldots + .125^2)}{6}}$$

$$= 1.163 \qquad\qquad (6\ \text{df})$$

which is remarkably close to the value of 1.075 found in the earlier section. The standard error is thus $s/\sqrt{2} = .822$, which is, of course, wider than in the replicated experiment because here we have only four observations at each level of each factor.

The important point is that, even in an unreplicated experiment, blocking will usually increase precision. It also improves accuracy (reliability of the conclusions) relative to those possible if all eight runs had been carried out on one car. Although it is not replication in the usual sense of repeating the entire experiment on a second car, one can see by inspection of Table 5 that each effect is estimated from the four observations in each block (since there are two observations at each level of each effect in each block). Thus we have replicated the contrast and have some idea of its consistency over two different cars.

An example of a very successful blocked fractional factorial experiment can be found in Young, Whitney and Abraham (1991).

How Randomization Protects Against Systematic Bias

During any experiment there is always some chance that misleading conclusions will result from an unexpected systematic shift in the process due to a change in one or more external (non-experimental) factors. The use of small blocks that can be completed in a relatively short period of time minimizes this danger since there will be relatively little chance for the non-experimental factors to change during the execution of the runs within the block. Block-to- block changes are estimated by the block term in the model (Model 3). As well, all effects are estimated independently in every block; if the order of the runs within a block is randomized separately for each block, there is virtually no chance that the effect of any factor will suffer the same degree of bias in all blocks. The combination of blocking and randomization will provide strong protection against bias.

Since none of the main effects or two-factor interactions were significant in this experiment, the observed difference between cars three and four can be used to simulate the effect of systematic bias occuring during an unblocked experiment. The standard order found in traditional contrast matrices or in Taguchi's orthogonal arrays always has one main effect contrast with the first half of the runs at its low level followed by the rest at its high level (Table 1). As a result, without randomization, any systematic shift in the process that occurs about half way through the experiment will result in this effect looking significant. A shift earlier or later may cause a two-factor interaction to appear significant. For instance, if the eight-run experiment described in the last section had been run in the order of Table 1 without being blocked, and the results of blocks three and four been obtained because of a change in car, operator, or batch of parts, then force would have looked highly significant as seen in the probability plot in Figure 1.

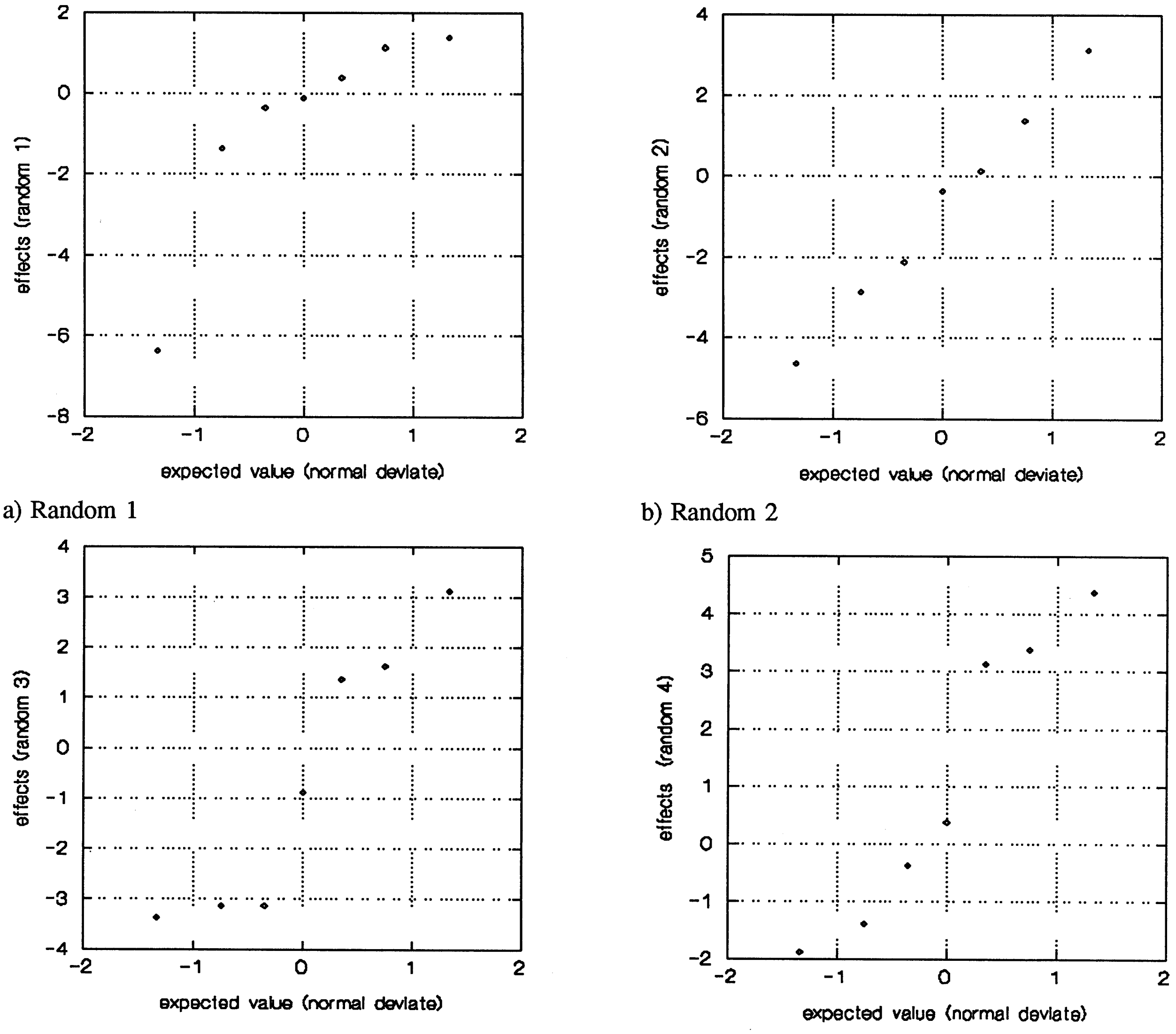
If a systematic shift in the process occurs at the beginning of the fifth run of an eight-run experiment, two of the $\binom{8}{4} = 70$ (eight choose four) possible arrangements of the signs of one contrast will link either the high ('+') or low ('-') level of the contrast with the last four runs. Since there are seven contrasts, a total of $7 \times 2/70$ or 20% of the possible randomizations would lead one of the seven contrasts being significant because of this shift. The other 80% of the time the bias would be spread more randomly over the seven contrasts. Although a 20% probability of a contrast looking significant for the wrong reason is rather bad, it is certainly better than the 100% probability when there is no randomization. A similar calculation for a sixteen-run experiment shows that only 0.23% of the time would a shift half way through the experiment correspond to either the high or low level of one of the fifteen contrasts. This is certainly a very good reason to use sixteen-run experiments if possible!

An indication of how randomization would help in an eight-run experiment can be obtained by ignoring the blocking, randomizing the order of the runs, and leaving the observed closing efforts in the same order. This simulates a major process shift after the fourth run in a situation in which none of the effects is significant.

A representative selection of four of one hundred such randomizations that were made is included in Table 5. Probability plots for these four are given in Figure 2.

Plots like that in part a) of Figure 2 occurred 24% of the time and corresponded to all four large observations coinciding with either the high or low levels of one of the seven contrasts. With a plot like this, most analysts would incorrectly decide that one effect is significant. Although this is not good, it is certainly better than not randomizing, which would result in one contrast always appearing significant.

Of more interest is the fact that a substantial number of plots looked like parts b) and c). Most experienced analysts would not judge any of the contrasts in these to be significant. Many, however, looked like part d) and would be harder to judge.



a) Random 1

b) Random 2

c) Random 3

d) Random 4

Figure 2: Probability plots of four randomizations of the eight-run experiment on cars three and four.

Clearly, in even an eight-run experiment, randomization offers some protection. There is, however a very good chance of drawing the wrong conclusions. This danger can, to a great extent, be avoided by blocking. The danger essentially disappears when a blocked sixteen-run experiment is carried out. Randomization of runs within blocks should still be used.

Conclusions

As far as the effect on closing effort is concerned, the influence of the stuffer is clearly negligible in comparison with other factors that vary from car to car. In view of this result, specifications were set up corresponding to the ranges studied in the experiment. The process engineers then shifted their emphasis to other more promising factors. Although the experiment did not solve the problem, it put to rest concerns regarding the stuffer. And it did this at a very low cost to the production department.

Blocking played a very major role in this experiment by greatly increasing the precision with which the influence of the stuffer could be estimated. The direct *replication* in blocks of the 2^3 runs adds greater credibility to the accuracy and generality of the conclusions drawn since the results were consistent over a representative sample of cars. It also provided an independent measure of experimental error, thus avoiding the judgment required by a probability plot analysis. *Randomization* of the order of the treatments provided insurance against unknown systematic bias leading to incorrect conclusions. Lack of care in choosing factors and levels and the absence of these three critical components of design are the most common cases of failed experiments.

It should also be emphasized that, although there is a greater risk of things going wrong in small eight- and nine-run experiments, it is still better to follow a systematic factorial plan that is randomized and blocked than to use the "lets try this" fire-fighting approach.

References

Box, G.E.P., Hunter, W.G. and Hunter, S. ***Statistics for Experimenters.*** John Wiley & Sons, New York, (1978).

Cox, D.R. ***Planning of Experiments.*** John Wiley & Sons, New York, (1958).

Daniel, C. and Woods, F.S. ***Fitting Equations to Data, 2nd Edition.*** John Wiley & Sons, New York, (1980).

Young, J.C. ***Experimental Designs: A Strategy for Implementation.*** Transactions of the 43rd ASQC Quality Congress. Toronto, Ontario, Canada, 34-40, (1989).

Young, J.C., Whitney, J.B. and Abraham, B. ***Design Implementation in a Foundry: A Case Study.*** Quality Engineering , Volume 3, (1991).