Quality Engineering, 4(3), 453–459 (1992)

GEORGE'S COLUMN

GEORGE E.P. BOX
Director of Research
Center for Quality and Productivity Improvement
University of Wisconsin—Madison
Madison, Wisconsin 53706

Teaching Engineers Experimental Design with a Paper Helicopter

When Søren Bisgaard, Conrad Fung, and I teach engineers about designed experiments, we find it very valuable to use a paper helicopter for illustration. We were introduced to this idea some years ago by Kip Rogers of Digital Equipment. Using the generic design shown in Figure 1 a "helicopter" can be made from an 8½ × 11 sheet of paper in a minute or so.

The scenario I'll describe requires three people who I'll call Tom, Dick, and Mary. To make an experimental run Tom stands on a ladder and drops the helicopter from a height of 12 feet or so while Dick times its fall with a stopwatch. We explain to the class that we would like to find an improved helicopter design which has a longer flight time. The helicopter can then be used to illustrate a number of important ideas.

Variation

We start with Tom dropping a helicopter made from blue paper. He drops it four times and we see that the results vary somewhat. This leads to a discussion of variation and to the introduction of the range and the standard deviation as measures of spread, and of the average as a measure of central tendency.

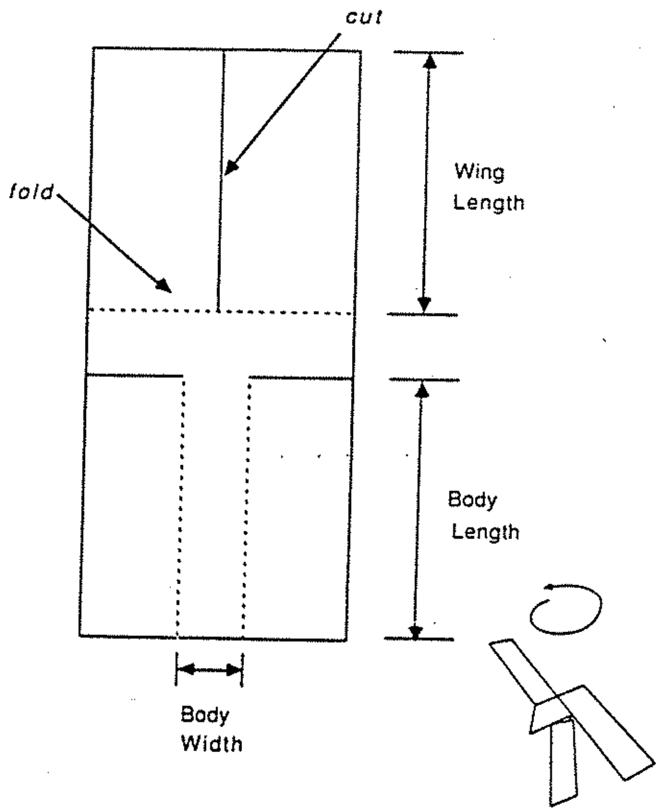

Figure 1. A paper helicopter.

Comparing Mean Flight Times

At this point Dick says "I don't think much of this helicopter design, I made this red helicopter yesterday and dropped it four times and I got an average flight time which was considerably longer than what we just got with the blue helicopter." So we put up the two sets of data for the four runs made with the blue helicopter and the four runs made with the red helicopter on the overhead projector and we show the two sets of averages and standard deviations. Eventually we demonstrate a simple test that shows that there is indeed a statistically significant difference in means in favor of the runs made with the red helicopter.

Validity of the Experiment

At this point Mary says "So the difference is statistically significant. So what? It doesn't necessarily mean it's because of the different helicopter design. The runs with the red helicopter were made yesterday when it was cold and wet, the runs with the blue helicopter were made today when it's warm and dry. Perhaps it's the temperature or the humidity that made the difference. What about the paper? Was the same kind of paper used to make the red helicopter as was used to make the blue one? Also, the blue helicopter was dropped by Tom and the red one by Dick. Perhaps they don't drop them the same way. And *where* did Dick drop his helicopter? I bet it was in the conference room, and I've noticed that in that particular room there is a draft which tends to make them fall toward the door. That could increase the flight time. Anyway, are you sure they dropped them from the same height?" So we ask the class if they think these criticisms have merit and most agree that they have, and they add a few more criticisms of their own. They may even tell us about the many uncomfortable hours they have spent sitting around a table with a number of (possibly highly prejudiced) persons arguing about the meaning of the results from a badly designed experiment.

We tell the class how Fisher once said of data like this that "nothing much can be gained from statistical analysis; about all you can do is to carry out a postmortem and decide what such an experiment died of." And how, some 70 years ago, this led him to the ideas of randomization and blocking which can provide data leading to unambiguous conclusions instead of an argument. We then discuss how these ideas can be used to compare the blue and the red helicopter by making a series of paired comparisons. Each pair (block) of experiments involves the dropping of the blue and red helicopters by the same person at the same location; you can decide which helicopter should be dropped first by, for example, tossing a penny. The conclusions are based on the *differences* in flight time within the pairs of runs made under identical conditions. We go on to explain however that different people and different locations could be used from *pair to pair* and how, if this were done "it would widen the inductive basis" as Fisher (3) said, for choosing one helicopter design over the other. If the red helicopter design appeared to be better, one would, for example, like to be able to say that it seemed to be consistently better no matter who dropped it or where it was dropped. As we might put it today, we would like the helicopter design to be "robust with respect to environmental factors such as the 'operator' dropping it and the location where it was dropped." This links up very nicely with later discussion of robust design of products.

A Fractional Factorial Design

Later on in the class, we use the paper helicopter to illustrate the running of a fractional factorial (orthogonal array) design. We suppose that a brainstorming session by an engineering design team on ways of improving the helicopter flight time has resulted in the selection of eight factors to be studied in a designed experiment. These selected factors are listed at the top of Figure 2 together with the two conditions (indicated by minus and plus signs) at which each will be tested. It is thought likely that only a few of these factors will have important large effects. We are thus in the familiar "Pareto" situation where, as Dr. Juran says we want to screen out "the vital few from the trivial many." The design, shown in Figure 2, is a fractional factorial design. Bisgaard (2) provides a very useful table of this and other eight and sixteen-run designs with a succinct description of their properties and analysis. Such designs, which were developed in England during and just after World War II, are particularly useful for this purpose of screening, and this one which is a 1/16th fraction of the full 2^8 (256 run) design has two very valuable properties (see e.g., Ref. 1).

1. If there are interactions between pairs of factors, they will not bias any of the 8 main effects of the factors.
2. If only up to 3 factors are of importance, the design will produce a complete 2^3 factorial design replicated twice in those three factors no matter which ones they are.

This latter property is particularly remarkable when we consider that there are 56 different ways of choosing 3 factors from 8. You can check it for yourself by picking any 3 columns in the design of Figure 2 and verifying that whichever 3

FACTORS		-	+	*effect*
Paper Type	(P)	Regular	Bond	0.13
Wing Length	(W)	3.00"	4.75"	0.77
Body Length	(L)	3.00"	4.75"	-0.40
Body Width	(B)	1.25"	2.00"	-0.02
Paper Clip	(C)	No	Yes	0.05
Fold	(F)	No	Yes	-0.10
Taped Body	(T)	No	Yes	-0.15
Taped Wing	(M)	No	Yes	0.17

Random Order	Standard Order	P	W	L	B	C	F	T	M	Flight Time
7	1	-	-	-	-	-	-	-	-	2.5
13	2	+	-	-	-	+	-	+	+	2.9
4	3	-	+	-	-	+	+	-	+	3.5
9	4	+	+	-	-	-	+	+	-	2.7
1	5	-	-	+	-	+	+	+	-	2.0
12	6	+	-	+	-	-	+	-	+	2.3
15	7	-	+	+	-	-	-	+	+	2.9
3	8	+	+	+	-	+	-	-	-	3.0
6	9	-	-	-	+	-	+	+	+	2.4
16	10	+	-	-	+	+	+	-	-	2.6
14	11	-	+	-	+	+	-	+	-	3.2
5	12	+	+	-	+	-	-	-	+	3.7
11	13	-	-	+	+	+	-	-	+	1.9
10	14	+	-	+	+	-	-	+	-	2.2
2	15	-	+	+	+	-	+	-	-	3.0
8	16	+	+	+	+	+	+	+	+	3.0

Figure 2. Results from 16 run fractional factorial experiments showing the factor levels and the calculated main effects of the eight factors.

you pick you have every combination of $(\pm, \pm, \pm)$ in these factors repeated twice over.

Flight times for the 16 helicopter types obtained from an experiment run in random order are shown in Figure 2. From these flight times, 8 main effects and 7 strings of two factorial interaction effects may be calculated.* These are plotted on probability paper in Figure 3 suggesting that real effects are associated with W (wing length) and, less certainly, L (body length). On the basis that the remaining effects falling around the straight line are mostly due to noise, we can summarize the data simply in terms of the inset diagram in Figure 3. Going back to the original data it will be seen, for example, that there are four runs with short wing length and short body length with flight times averaging 2.6 seconds and 4 runs with long wing length and short body length averaging 3.3 seconds and so on. These averages are set out at the corners of the square in the inset diagram. A direction in which one might expect still longer flight times by using larger wings

*It is supposed in this analysis that interactions between 3 or more factors can be ignored. A fuller discussion of such analyses can be found, e.g., Ref. 1, p. 402.

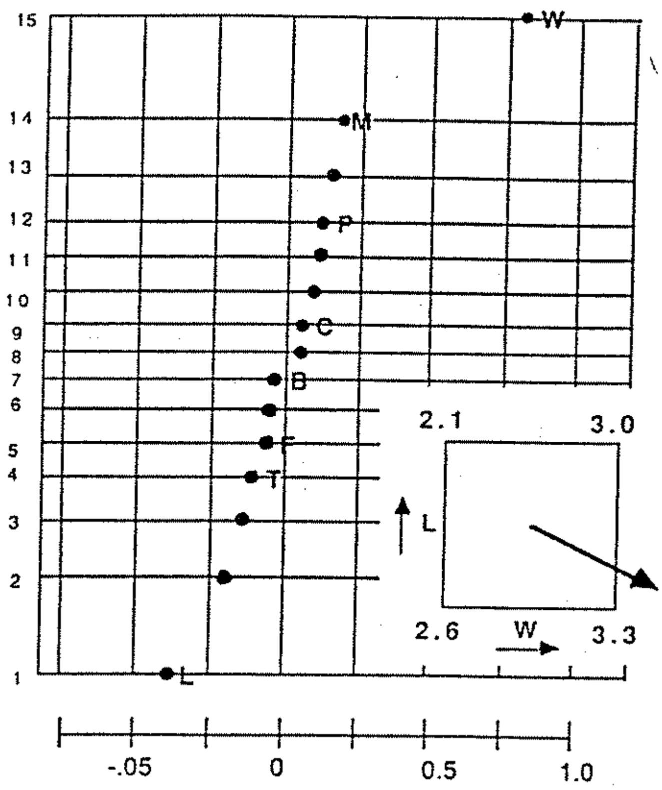

Figure 3. A normal plot of the effects from the helicopter experiment. The inset diagram summarizes the conclusions.

with a shorter body is indicated by the arrow. Thus the experiment immediately provides not only an improved helicopter design but also indicates the direction in which further experimentation should be carried out and so demonstrates the value of the sequential approach to experimentation—learning as you go. Another aspect of this approach is highlighted by discussing with the class whether they are satisfied with *flight time* as the sole criterion. In earlier lectures we have emphasized to the class that what happens in each run of an experiment must be carefully documented—for example, the fact that helicopter #7 hit the table leg and that that run had to be repeated. Such careful observation might suggest, for example, that an additional criterion that should be included in future experimentation, is flight *stability*. This teaches the lesson that the *criteria* to be used in assessing the results may need to be modified or totally changed during an investigation as we learn more of the phenomena under study. Appropriate and feasible objectives *cannot* always be determined in advance.

Management of Experimentation

In running an experiment as complex as this, the safest assumption is that, unless extraordinary precautions are taken, it will be run incorrectly. Therefore, the opportunity should be taken to involve the class in the careful organization of the experiment. In particular, members of the class should be assigned to systematically check, and recheck independently, that each of the 16 helicopter designs to be flown corresponds exactly to the specification set out in the appropriate row of Figure 2. Our course for engineers lasts only a few days so we find it necessary to prepare the paper helicopters in advance. After the preliminary explanation and the careful checking, the actual running of the experiment takes less than 6 minutes.

No elaborate analysis is needed for two-level experiments of this kind and certainly no analysis of variance table, which at this stage and for this purpose serves only to waste time and confuse the class. In earlier discussions, members of the class have already satisfied themselves, by one or two hand calculations, that factorial effects are just the differences between the average results at the plus and minus levels of a given factor. Also the rationale of Daniel's normal plot has already been explained. So for the helicopter experiment we enter the data in the computer as it becomes available and use the SCA program (4) to calculate the effects at once, and to produce the normal plot which is immediately projected onto the overhead screen.

We find that participatory demonstrations of this kind even with such a simple device as a paper helicopter seizes the imagination of the engineer and produces very rapid learning.

Acknowledgments

This research was sponsored by the National Science Foundation under Grant No. DDM-8808138, and by the Vilas Trust of the University of Wisconsin, Madison.

References

1. Box, G. E. P., Hunter, W. G., and Hunter, J. S., *Statistics for Experimenters*, John Wiley, New York 1978, p. 398.
2. Bisgaard, S., *A Practical Aid for Experimenters*, Starlight Press, Madison, WI, 1988.
3. Fisher, R. A., *The Design of Experiments*, Oliver and Boyd, Edinburgh and London, 1935.
4. Scientific Computing Associates, P.O. Box 625, Dekalb, IL 60115.