TREATMENT DESIGN AND THE INTERPRETATION OF EXPERIMENTAL RESULTS[1/]

Walter T. Federer, Cornell University

The planning of an experiment is one of the most important if not the most important phase in the successful and efficient conduct of an experiment. This important part of experimentation does not receive the attention that it should. One phase in planning experiments is the selection of treatments or the "treatment design." On the average, it could be said that treatment design is as important as the combined effect of all other factors related to the successful completion of an experiment. This includes the selection of an experimental design, size and shape of experimental unit, number of replicates, control of biases, summary and interpretation of experimental results, etc. (3, Ch. I). The comparisons that can and should be made are determined to a large extent by the treatments included in the experiment. The treatment design determines whether or not evidence on specific questions can be obtained.

A set of selected examples will be used to illustrate the importance of treatment design and the questions on which evidence is available given a specific set of treatments. The treatment design for each experiment must be considered individually and in light of the hypotheses to be tested.

[1/] Paper No. 357 Department of Plant Breeding and paper No. BU-45 Biometrics Unit.

The Examples

Example 1. Nine fertilizer treatments were compared in a randomized complete block experiment on tomatoes using six replicates. One of the treatments was a check representing no fertilizer whatsoever. The other eight treatments represented a 2^3 factorial of the three factors: nitrogen, phosphorous, and potash. The plots receiving the eight treatments in the 2^3 factorial arrangement were treated with a uniform application of lime. Thus, the nine treatment combinations, using standard notation (3,7) were: 0000, 1000, 1010, 1100, 1110, 1001, 1011, 1101, 1111.

The experimenter who conducted this experiment summarized the yield data as follows. First, he computed an F test and found the observed F to be less than the tabulated F value for 8 and 40 degrees at the 5 per cent level. Secondly, he compared each of the eight treatments of the 2^3 factorial with the check using a t test and found that each of the eight comparisons exceeded the tabulated t value for 40 degrees of freedom at the one per cent level. Thirdly, he concluded that the F test was not a discriminating statistic and suggested that experimenters use the t test in preference to the F test.

Apparently this experimenter did not know the relationship between t and F tests or the types of and relations among hypotheses compared by the tests performed. Aside from these facts, there are more basic questions which we as statisticians should ask about experiments of this type; these are: (i) Why were these particular nine treatment combinations selected? (ii) Were the questions on which answers were sought, clearly formulated prior to the conduct of the experiment? (iii) Is there other evidence which

would indicate the relative performances among these nine treatments? (iv) Assuming that it was desired to describe the response surface generated by the 2^3 treatments, why not compute main effect and interaction mean squares?

In this particular experiment, it was found that the comparison of the mean of treatment 0000 with the mean of the other eight yielded an F which exceeded the tabulated F value at the one per cent level. None of the main effect or interaction mean squares were significantly larger (five per cent level) than the error mean square. In fact, the mean square for the comparison among the eight treatments in the 2^3 factorial was approximately equal to the error mean square. Thus, from this analysis we could conclude that the eight treatments in the 2^3 factorial could be considered as forming a group said to be non-heterogeneous and that the mean of check differed from the mean of the treatments in the 2^3 factorial. This summarizes all of the information in the experiment. Why go further? The particular nine treatments selected in this experiment indicate that the experimenter was interested in the above comparisons. There is no need to make any other comparisons since the eight individual degrees of freedom summarize the whole of the information for the nine treatments in this experiment.

Example 2. Let us assume that the experimenter wishes to describe a response surface of unknown form for a given set of factors. This can be accomplished by selecting a factorial arrangement of a number of levels of the factors affecting the phenomenon under consideration. Once the factorial set of treatments has been selected and the experimenter has stated that he is interested in main effects and interactions, the form of the analysis is

the standard one for factorial experiments.

Occasionally an experimenter selects a factorial arrangement of treatments and states that he is interested in main effects and interactions; then, he uses a multiple range test or a multiple comparisons test (3, Ch. II) to compare pairs or means. Such a procedure negates the original assertion by the experimenter that he is interested in main effects and interactions. In most situations such procedures disagree with the comparisons dictated by the treatment design and should not be used.

There is, however, a situation in which the procedure of applying a multiple range test, a multiple comparisons test, or a multiple F test would be considered in agreement with the comparisons dictated by the treatment design. If the experimenter states that interactions are non-existent or negligible and that the interest lies in comparisons among pairs and groups of the main effect means, then a multiple range test procedure would be appropriate.

Pairwise comparisons of means in a factorial experiment suggests that the experimenter is searching for an optimum combination of factors and is interested in selecting one treatment combination from the entire range of treatment combinations. If such is the case, the experimenter should use one of the treatment designs suggested by Box et al. (1,2), the procedure described by Friedman and Savage (4), or the procedure described by Kiefer and Wolfowitz (8). These procedures are designed specifically for estimating a unique maximum on the response surface. The factorial treatment design is designed to describe the entire response surface for the particular levels and factors included in the experiment. Precise statements of objectives of

an experiment are mandatory if the correct treatment design is to be selected. Vagueness concerning objectives may, and often do, lead to the use of incorrect procedures and incorrect treatment designs.

Example 3. The following data were obtained from Harter and Lum (6):

	Column			
Row	1	2	3	Total
1	2	4	12	18
2	4	8	24	36
Total	6	12	36	54

The analysis of variance for the above data are:

Source of variation	df	Sum of squares
Row	1	54
Column	2	252
Interaction	2	28
Non-additivity	1	28
Residual	1	0

where the sum of squares for non-additivity is calculated as described by Tukey (9). If the above data are transformed to logarithms, the resulting analysis of variance on the logarithms of the original data are:

Source of variation	df	Sum of squares
Row	1	.1359
Column	2	.6157
Interaction	2	0

Thus, we see that a simple transformation of the data removes the interaction. Tukey's one degree of freedom for non-additivity is useful in removing interactions in the original data whenever the transformation is of the form x^p or $(x+C)^p$. The logarithmic transformation fits into this group because of the analogy between log x and x^0 (6,9). Therefore, for transformations in this group Tukey's one degree of freedom for non-additivity effectively does the same thing as a transformation.

In analyzing data from factorial experiments, it is desirable to remove the part of the interaction that is removable by a transformation in order to make the data more nearly additive. If the experimenter removes removable interactions whenever possible, the number of cases in which there are significant interactions will decrease considerably. Likewise, an experimenter should not select a transformation which produces a spurious interaction. These items should be borne in mind in selecting treatments in factorial experiments.

Example 4. Whether a factorial arrangement of treatments or some sequential procedure to estimate an optimum (1,2,4,8) is used, it is highly questionable if the present method of adding fixed amounts of factors to growing organisms yields answers to the desired questions. A procedure that is more realistic in a fertilizer trial, for example, is to determine the actual amounts of the various elements of factors in plants. Then, the levels of the elements of the factors in the plant would determine the amount of the factor applied. Regardless of the amount of a factor in the soil, the conditions may be such that the plant cannot utilize the factor or the elements in it. Therefore, the amount of the elements of the factor in the plant and

not the amount applied or in the soil determines the level of the elements of the factor in the plant.

In light of the above, it would be more realistic in describing the entire response surface or in describing one point on the response surface such as a maximum, to set up the factorial or sequential experiment by determining the actual level of the elements of the factor in the plant; then, a sufficient amount of the factor would be added to bring the amount in the plant up to the desired level. Such an experiment would yield answers on actual optimums and on the relationships of levels of factors on the plant character, say yield, under study. Even where crop-logging techniques are fairly advanced (e.g. in Hawaii on sugar cane and pineapple), no experiments of this type have been conducted. The reason may be that the experimenters have not precisely and clearly formulated the questions on which answers are sought or if they have, this step in the planning of experiments was not coordinated with the related step of selecting the treatments to be included in the experiment.

Example 5. In certain factorial experiments the treatments are such that the standard method of analysis for a factorial experiment indicates an interaction when actually none is present. For example, suppose that a 2x3 factorial arrangement consists of two methods of application (m_0 and m_1) and three levels of a fertilizer (f_0, f_1 and f_2) with the lowest level being zero (no fertilizer), that there is a difference in the yields obtained from the two methods of application and from the fertilizers, and that the results in figure 1 were obtained from the experiment. The standard analysis would indicate an interaction. A more appropriate analysis assuming that the methods

of application in the absence of a fertilizer have no effect, is (r replicates):

Source of variation	df	
Treatments	4	
No fertilizer plots vs. rest	1	
Level f_1 vs. Level f_2 = F	1	
Method m_0 vs. Method m_1 = M	1	
F x M	1	
Duplicate plots of no fertilizer plots	r-1	error
Interaction	5(r-1)	error

where the interaction (FxM), F, and M are computed from the four treatments m_0f_1, m_0f_2, m_1f_1, and m_1f_2. From figure 1, it is evident that this interaction has zero sum of squares, and hence contradicts what was obtained from the standard analysis.

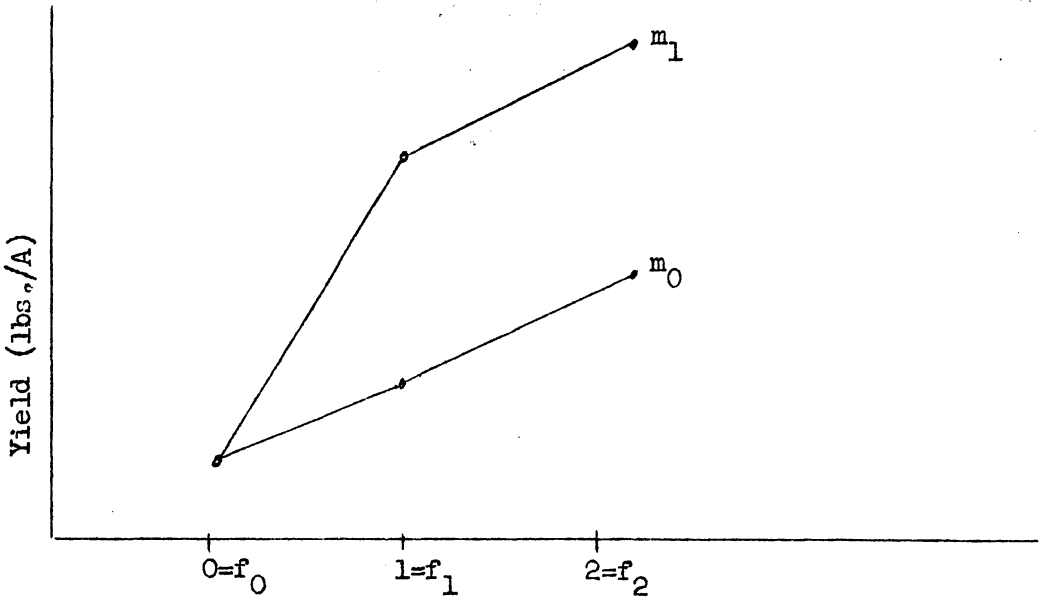

Figure 1. Yields for two methods of application and three levels of a fertilizer.

A second example illustrating the same thing is found for material which is not available at the same time. For example, the character might be weight of sugar cane plants below the 10th internode or amount of nitrogen, potash, or phosphorous in the part of the plant between the 8th and 10th internodes in sugar cane. If age of plant is one variable and amount of nitrogen applied is the other, it is possible to obtain an age x level of nitrogen interaction simply by starting too early. An age x nitrogen interaction would be obtained if analysis of the data started at 3, 4, or possibly 5 months of age; no interaction is obtained if the analysis starts at 6 months of age.

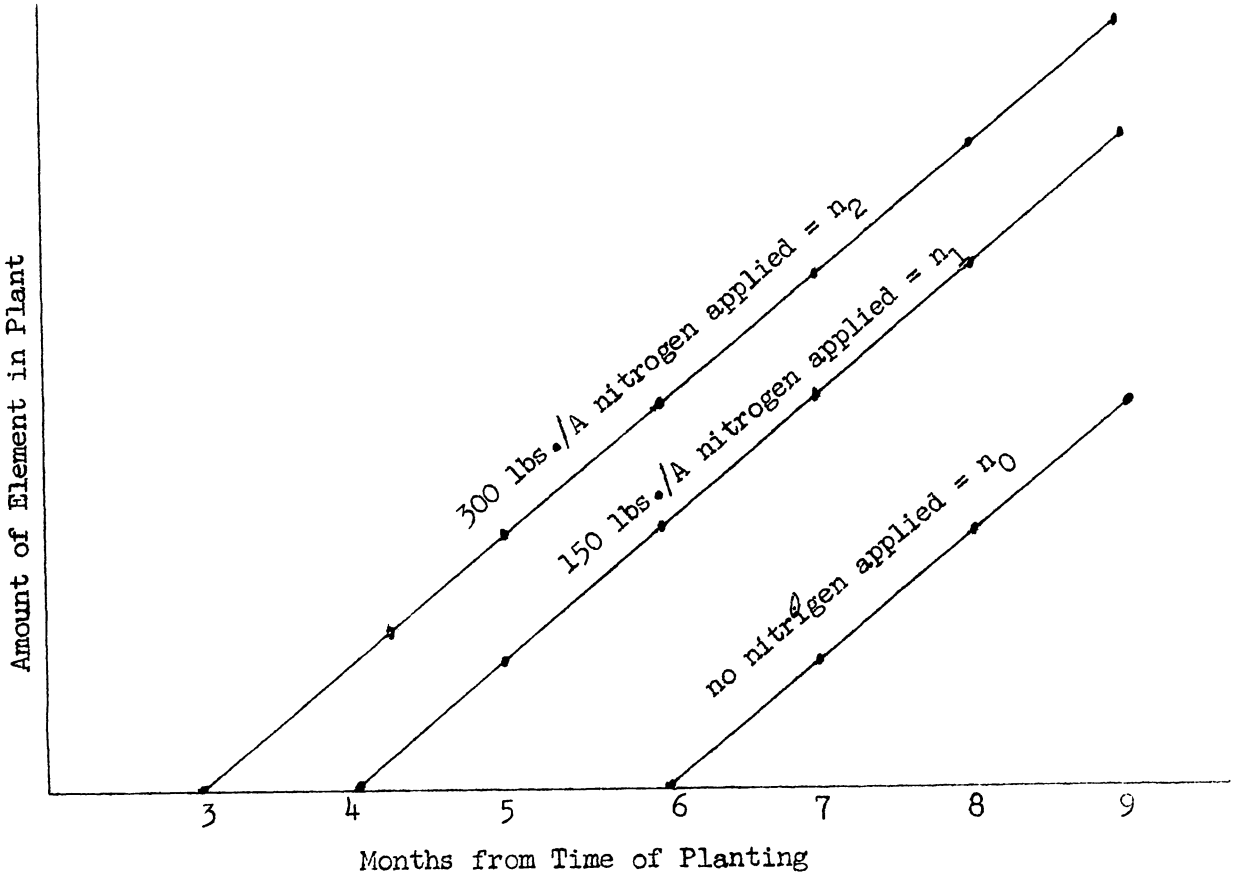

Figure 2. Yields for three amounts of nitrogen applied to the plots for different ages of plants.

Examples like the above indicate a type of spurious interaction that arises from faulty or incomplete analysis. We know before running the experiment that a "significant interaction" will be obtained if data are not available for all treatments in the experiment.

Example 6. The kind of treatments in an experiment is determined by the phase of experimentation which in turn should determine the error rate used in analyses (5,10). For example, in a plant breeding experiment, the plant breeder starts with a large number of entries in the first step of an evaluation program of a lot of genetic material. The first step may be to discard all entries with undesirable agronomic traits except for yield. The second and later steps may involve yield trials on the surviving members at the end of each step. Finney (3a) has proposed a procedure for determining the number of survivors to be tested at the various stages. If Finney's procedure is not appropriate the breeder may wish to change the error rate per experiment in calculating tests of significance and multiple range tests (5,10). It may be appropriate to use Type I and II errors of 25%, say, in the preliminary phases whereas Type I and Type II errors of one to five per cent, or lower, may be desired in the final phases of an experimental program. Thus, the type of treatments in an experiment, or the treatment design, may, and probably should, determine the error rates selected for summarizing the data.

Example 7. The genetic treatment design for plants, say, involving two parents (P_1 and P_2), the cross between the parents ($P_1 \times P_2 = F_1$), the second generation from the selfed hybrid ($F_1 \times F_1 = F_2$), and the backcross to the two

parents ($B_1=F_1xP_1$ and $B_2=F_1xP_2$) is suitable only if reciprocal crosses are identical. If, however, one cannot positively state that reciprocal crosses are identical the treatment design must contain P_1, P_2, P_1xP_2, P_2xP_1, F_1xF_1, F_1xP_1, P_1xF_1, F_1xP_2, and P_2xF_1 if valid inferences are to be drawn from the data. The treatment design determines the type of inferences that can validly be drawn from the data.

For certain genetic inferences it is necessary to have additional backcrosses and selfed filial generations. It is not uncommon to find the genetic treatment design of P_1, P_2, F_1, F_2, B_1 and B_2 and then to find that the experimenter attempts to make inferences which require the F_3, F_4 and F_5 generations, say, and the comparable possible backcrosses. Again, the principle that the treatment design is a most important part of experimentation is enunciated.

Example 8. In certain experiments on fungicides, herbicides, or insecticides the success of an experiment is entirely dependent upon the disease organism, the weed, or the insect being present in the experimental area if differences between treatments are to be evaluated. The treatment design may be perfect but the experimental conditions must be such that differences between treated and untreated plots can be evaluated. For example, soil fumigation tests on sugar cane are of no value on highly productive areas; the tests should be placed on "sick" or low producing areas. Even with such a restriction not all tests will be successful unless the organism or soil condition affected by soil fumigation is present in all low producing areas. Likewise, yield tests on varieties or fertilizers located in the Sahara Desert are worthless because the experimental conditions are such that yield differences cannot be evaluated.

The treatment design must be such that the value of any experiment in evaluating effects can be assessed. Untreated plots must be included to determine the level of insect infestation, the number of weeds, etc. that are present in the experimental area. If no insects, weeds, etc. are present the experiment turns out to be just an observation.

Discussion and Summary

The examples presented indicate the importance of treatment design, or the selection of treatments, in experimental work. Although the examples do not cover the entire area of experimentation and do not bring to light all of the problems involved, they are of sufficient diversity to indicate the types of problems encountered in treatment design. The relationship of treatment design and comparisons made among the treatments is probably of most concern. However, treatment design and error rates in experimentation, treatment design and experimental conditions, and treatment design and inferences drawn from the data also require close attention by the experimenter and statistician alike.

All steps in the planning and conduct of an experiment are important but the treatment design is considered to be the most important single factor in the successful and efficient completion of an experiment. In fact, treatment design over all types of experiments may be as important as all other factors combined.

1. Box G. E. P. The exploration and exploitation of response surfaces: some general considerations and examples. Biometrics 10:16-60, 1954.

2. Box, G. E. P. and Hunter, J. S. Multifactor experimental designs. Ann. Math. Stat. 28:195-241, 1957.

3. Federer, W. T. Experimental Design. Macmillan, N. Y., 1955.

3a. Finney, D. J. Statistical problems of plant selection. Almqvist and Wiksells Boktryckeri AB, Uppsala, 1957.

4. Friedman, M. and Savage, L. J. Planning experiments seeking maxima. Ch. 13 in Techniques of statistical analysis. McGraw-Hill, N. Y., 1947.

5. Harter, H. L. Error rates and sample sizes for range tests in multiple comparisons. Biometrics 13:511-536, 1957.

6. Harter, H. L. and Lum, Mary D. A note on Tukey's one degree of freedom for non-additivity. Unpublished paper presented at annual meeting of Amer. Stat. Assoc. and the Biometric Soc., 9/10-13/57.

7. Kempthorne, O. The design and analysis of experiments. Wiley, N. Y., 1952.

8. Kiefer, J. and Wolfowitz, J. Stochastic estimation of the maximum of a regression function. Ann. Math. Stat. 23:462-466, 1952.

9. Tukey, J. One degree of freedom for non-additivity. Biometrics 5:232-242, 1949.

10. Tukey, J. The problem of multiple comparisons. Ditto, Princeton University, 396 pp., 1953.