# A FRESH LOOK AT THE BASIC PRINCIPLES OF THE DESIGN AND ANALYSIS OF EXPERIMENTS

F. YATES
ROTHAMSTED EXPERIMENTAL STATION, HARPENDEN

## 1. Introduction

When Professor Neyman invited me to attend the Fifth Berkeley Symposium, and give a paper on the basic principles of the design and analysis of experiments, I was a little hesitant. I felt certain that all those here must be thoroughly conversant with these basic principles, and that to mull over them again would be of little interest.

This, however, is the first symposium to be held since Sir Ronald Fisher's death, and it does therefore seem apposite that a paper discussing some aspect of his work should be given. If so, what could be better than the design and analysis of experiments, which in its modern form he created?

I do not propose today to give a history of the development of the subject. This I did in a paper presented in 1963 to the Seventh International Biometrics Congress [14]. Instead I want to take a fresh look at the logical principles Fisher laid down, and the action that flows from them; also briefly to consider certain modern trends, and see how far they are really of value.

## 2. General principles

Fisher, in his first formal exposition of experimental design [4] laid down three basic principles: replication; randomization; local control.

Replication and local control (for example, arrangement in blocks or rows and columns of a square) were not new, but the idea of assigning the treatments at random (subject to the restrictions imposed by the local control) was novel, and proved to be a most fruitful contribution. Its essential function was to provide a sound basis for the assumption (which is always implied in one form or another) that the deviations used for the estimation of error are independent and contain all those components of error to which the treatment effects are subject, and only those components. When a randomized design is used and correctly analyzed disturbances such as those arising from real or imagined fertility gradients in agricultural field trials, and the fact that neighboring plots are likely to be more similar than widely separated plots, can be ignored in the interpretation of the results. The results (yields, and so forth) can in fact be treated as if they were normally and independently distributed about "true" values given by additive constants representing the treatments and block or other local control effects. Valid estimates of the treatment effects and their errors can then be obtained by the classical method of least squares. The analysis of variance (another of Fisher's brilliant contributions) formalizes the arithmetic of this procedure, and permits its extension to more complicated designs, such as split plots, involving a hierarchy of errors.

There is, of course, nothing sacrosanct about the assumptions of normality and additivity, and alternative assumptions can be made if these appear appropriate. What is not sometimes recognized, however, is that the results of one small experiment provide very weak evidence on which to base alternative assumptions. The practical experimenter, or the statistician who works for him, bases his assumptions on long experience of the behavior of the type of material he is handling, and has devices, such as transformations, which enable him to reduce his data to a form which, he is reasonably confident, permits him to apply standard methods of analysis without serious danger of distortion.

This point, I think, Fisher never sufficiently emphasized, particularly as he frequently emphasized the opposing point that each experiment should be permitted to determine its own error, and that no *a priori* information on error should be taken into account.

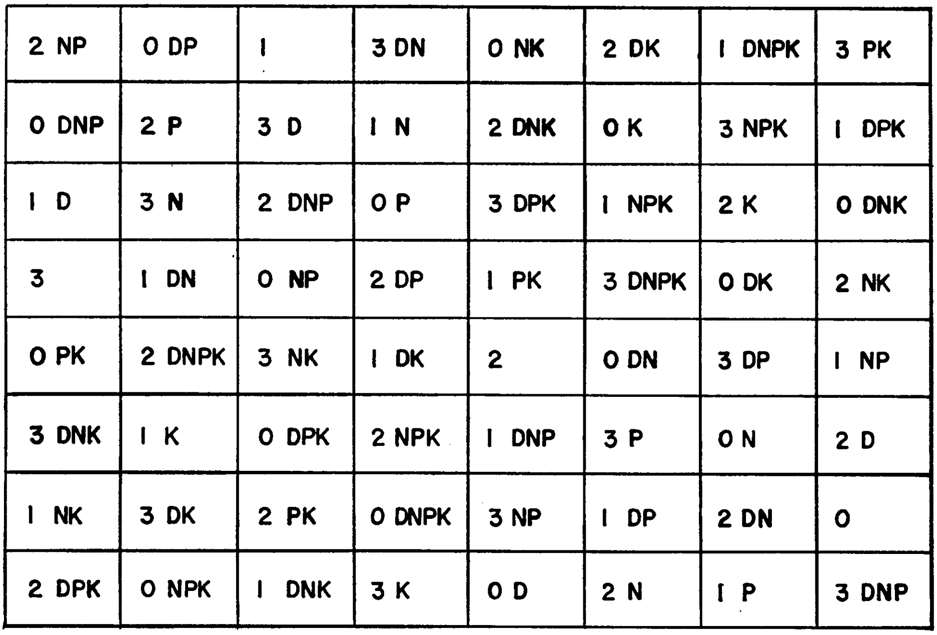
Nevertheless, Fisher, with his sound practical sense, drew the line between *a priori* and current information where, in the material he was handling, it should be drawn. In other circumstances he would undoubtedly have approved of other methods. When, for example, laboratory determinations of the content of a particular chemical compound are being made on a series of substances, with duplicate determinations on each, that is, 1 d.f. for error for each experiment, determination of error from a control chart, analogous to that used in quality control, is clearly preferable to treating each experiment as an independent entity.

Two points are important here. First, the experimenter does not want to be involved in a haze of indecision on the appropriate methods to apply to the analysis of each particular experiment. Second, when considering the results of an experiment, he should have clearly segregated in his mind the information on the treatment effects provided by the current experiment and that provided by previous experiments with the same or similar treatments.

## 3. Some points on randomization

There is one point concerning randomization in experiments to which Fisher always appeared to turn a blind eye. As soon as, by some appropriate random process, an experimental layout is determined, the actual layout is known, and can be treated as supplementary information in the subsequent analysis if this appears relevant. Usually, and rightly, the experimenter ignores this information, but there are occasions when it appears wrong to neglect it. Thus, in an agricultural experiment the results may indicate a fertility gradient which is very imperfectly eliminated by the blocks. Should the experimenter not then be permitted to attempt better elimination of this gradient by use of a linear regression on plot position, which can very easily be done by a standard covariance analysis? My own opinion is that when a large and obvious effect of this type is noticed, a statistician would be failing in his duty if he did not do what can be done to eliminate it. But such "doctoring" of the results should be the exception rather than the rule, and when it is resorted to, this should be clearly stated.

Such operations will in general introduce bias into the estimate of error, in the sense that for a fixed set of yields, the average error mean square, for all admissible randomization patterns of dummy treatments, will no longer equal the average treatment mean square. Absence of such bias has been used, by Fisher and others, as one of the justifications for randomization, and as a criterion for the validity of specific randomization procedures. The condition is certainly necessary to ensure full validity of the $t$ and $F$ tests, but can scarcely be regarded as sufficient to ensure a good experimental design. An example is provided by quasi-Latin squares [13]. In this type of design confounding of one set of interactions with the rows of a square, and another set with the columns, enables a $2^6$ design, for example, to be arranged in an 8 × 8 square of plots, row and column differences being eliminated as in a Latin square. The design



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2 NP | 0 DP | 1 | 3 DN | 0 NK | 2 DK | 1 DNPK | 3 PK |
| 0 DNP | 2 P | 3 D | 1 N | 2 DNK | 0 K | 3 NPK | 1 DPK |
| 1 D | 3 N | 2 DNP | 0 P | 3 DPK | 1 NPK | 2 K | 0 DNK |
| 3 | 1 DN | 0 NP | 2 DP | 1 PK | 3 DNPK | 0 DK | 2 NK |
| 0 PK | 2 DNPK | 3 NK | 1 DK | 2 | 0 DN | 3 DP | 1 NP |
| 3 DNK | 1 K | 0 DPK | 2 NPK | 1 DNP | 3 P | 0 N | 2 D |
| 1 NK | 3 DK | 2 PK | 0 DNPK | 3 NP | 1 DP | 2 DN | 0 |
| 2 DPK | 0 NPK | 1 DNK | 3 K | 0 D | 2 N | 1 P | 3 DNP |

FIGURE 1

$4 \times 2^4$ factorial design in an $8 \times 8$ quasi-Latin square.

is undoubtedly useful, but can with some frequency give an experimental layout in which the contrast between the two pairs of diagonally opposite 4 × 4 squares represents the main effect of one factor. This I first noticed when exhibiting in 1945 a slide of a $4 \times 2^4$ experiment done at Rothamsted in 1939 (figure 1) which had previously been exhibited in 1940 without exciting any comment! This unfortunate contingency is due to the fact that the required confounding is only possible with one rather special basic square. The defect can for the most part be obviated by what I have termed restricted randomization [6]. By excluding the most extreme arrangements, both those of the type mentioned and the complementary type which may be expected on average to be particularly accurate, the unbiased property of the error is preserved.

This throws light on the classical problem of the Knut-Vik or Knight's Move 5 × 5 Latin square, typified by figure 2(a). This was strongly advocated by some as likely to be more accurate than a random Latin square, and Tedin [10] showed by tests on uniformity trials that this was indeed so. The estimate of error, however, is necessarily biased in the opposite direction, so that the results appear *less* accurate than those of a random square. But, of course, a Knut-Vik square *might* be obtained by randomization. If this happens, should the experimenter reject it, and rerandomize? If so, the unbiased property of error will not hold over all remaining squares. This dilemma can in fact be neatly overcome by also excluding the diagonal squares, typified by figure 2(b), which Tedin also investigated and showed to be less accurate than the "random" squares he tested, though because of a small arithmetical error he did not recognize that the loss of accuracy in these squares exactly equals the gain in accuracy in the Knut-Vik

(a)

| | | | | |
|---|---|---|---|---|
| A | B | C | D | E |
| D | E | A | B | C |
| B | C | D | E | A |
| E | A | B | C | D |
| C | D | E | A | B |

(b)

| | | | | |
|---|---|---|---|---|
| A | B | C | D | E |
| E | A | B | C | D |
| D | E | A | B | C |
| C | D | E | A | B |
| B | C | D | E | A |

FIGURE 2

Knut-Vik and diagonal 5 × 5 Latin squares.
(a) Knut-Vik square. (b) Diagonal Latin square.

squares. This follows immediately from the fact that the four sets of 4 d.f. given by the treatment contrasts of the two Knut-Vik squares and the two diagonal squares are mutually orthogonal, and therefore together comprise the 16 d.f. left after eliminating rows and columns.

From the practical point of view none of this is of great importance, except in such special types of design as quasi-Latin squares. If the experimenter rejects arrangements with obvious systematic features, such as a diagonal pattern in a Latin square, he will not appreciably bias the estimate of error, as their chance of occurrence is very small (1/672 for a $5 \times 5$ diagonal square, for example). However, the exclusion of the complementary arrangements by formal application of restricted randomization is possibly worth while, not because it eliminates any general bias over all admitted arrangements, but because it eliminates arrangements that are particularly likely to give an overestimate of error. It should not be forgotten that the experimenter is much more concerned with the trustworthiness of estimates obtained from the arrangement actually selected, than with the behavior of these estimates in a hypothetical population of all admissible arrangements.

It may be asked, if the Knut-Vik squares are known to be on average more accurate, why not always use these, and accept an overestimate of error? There are two objections to this. First, valid estimates of error are in fact often required in experimental work, not only as a basis for tests of significance and fiducial limits in individual experiments, but also for investigating secondary points, for example, variation in treatment effects, over a set of experiments. Second, randomization not only provides valid estimates of error, it also eliminates distortions, which can be large, because, for example, one treatment always occurs in a fixed relation to another. With a randomized design the experimenter can examine his results objectively without continually looking over his shoulder to see if the apparent conclusions require qualification because of some statistical oddity in the design.

## 4. Estimation and tests of significance

Fisher, I think, tended to lay undue emphasis on the importance of formal tests of significance in experimental work. Many experiments have as their main object the estimation of effects of one kind or another. Often it is well known before the experiment is started that the treatments tested will have some effect. What is then required is an efficient estimate of these effects and a valid and reasonably accurate estimate of their error.

In part this emphasis on tests of significance is attributable to the way in which the subject developed, and to the fact that in the simpler types of experiment the treatment means furnish efficient estimates, whereas the correct estimation of error requires more subtle theory and more extensive computation; in part to the demands of experimenters, particularly biologists, to many of whom the attainment of a significant result seemed more or less equivalent to a new scientific discovery. Fisher himself was also much concerned with the logic of inductive inference, in which tests of significance play a central part.

The emphasis on tests of significance has undoubtedly had unfortunate consequences, both at the practical level, and in theoretical work. Too much effort has been devoted to the investigation of minor points of little real importance. This has resulted in proliferation of alternative methods of analysis, hedged about with restrictions and qualifications, to the confusion of the practical worker.

There is a logical point of some importance concerning alternative tests of significance on the same material. Although with large samples alternative tests which have similar power functions may be expected to give similar results when applied to a given set of data, provided these data conform to the basic assumptions on which the tests are based, this is by no means so with small samples. Consequently two statisticians applying two different tests, both of which are "reasonable," may arrive at very different conclusions. This situation is, to say the least, unfortunate.

An example is provided by the randomization test in experimental design. Fisher originally gave an example of this test in *The Design of Experiments* ([5], 1935) to provide confirmatory evidence of the validity of the $t$ test on the type of data to which it is usually applied without hesitation by the practical statistician. Unfortunately, this was taken to imply that the randomization test, because it made fewer assumptions, was somehow better, and that if the two tests did not agree on a particular set of data, the $t$ test was incorrect. Following this line of thought, Welch [11] evolved a method of "correcting" $F$ tests in randomized block and Latin squares so as to conform approximately to randomization tests.

Fisher did not regard the regular use of randomization and other nonparametric tests as reasonable. As he wrote in the second edition of *The Design of Experiments* ([5], 1937), "they were in no sense put forward to supersede the common and expeditious tests based on the Gaussian theory of errors." He did not, however, ever seriously discuss the question of what should be done when alternative tests give different verdicts, and in various passages in *Statistical Methods for Research Workers* [4], which were never amended, encouraged the statistician to look around for the test giving the highest significance. This is a pity.

## 5. Fixed, random, and mixed effects models

The analysis of variance, in the form originally proposed by Fisher, and developed by him and his coworkers, rapidly became the accepted method of analyzing replicated experiments. Once the requirements of orthogonality were understood it was successfully applied to very complex types of experiment, for example, those involving a hierarchy of split plots, partial or total confounding and fractional replication.

In addition to providing estimates of error and tests of significance for the various classes of effect, the results of an analysis of variance of experimental data can, if required, be used to estimate the variance components attributable to different classes of effect. Indeed in *Statistical Methods for Research Workers* ([4], 1925) the reader is first introduced to the analysis of variance in this context, as an alternative to intraclass correlation; this, as Fisher said, was "a very great simplification."

For most experiments the estimation of variance components is irrelevant, but such estimates are sometimes required. When, for example, in plant breeding work a random sample of varieties is selected for test, the varietal component of variance may be of interest. Similarly, when a fertilizer is tested on several fields selected at random, the component of variance of the response will represent the true variation in response to the fertilizer (apart from year to year variation).

All this was well known before the war and accepted by those using the analysis of variance to interpret experimental results. Unfortunately, after the war a new concept of fixed and random effect models was introduced. The trouble appears to have started with a paper by Eisenhart [2], which discussed the assumptions underlying the analysis of variance. In the course of this discussion he distinguished what he termed "Model I" or the *fixed effects model* and "Model II" or the *random effects model.* Although there is no real difference in his treatment of these two models, different symbols are used and equivalent formulae for expectations of mean squares in consequence look different. Furthermore, Eisenhart appeared to think that there was a genuine difference between them, or at least encouraged his readers to believe this. He wrote:

"*Which Model—Model I or Model II?* In practical work a question that often arises is: which model is appropriate in the present instance—Model I or Model II? Basically, of course, the answer is clear as soon as a decision is reached on whether the parameters of interest specify *fixed relations*, or *components of random variation.*"

Be that as it may, the hare, once started, could not be stopped. Differences in formulae, arising from differences in definition, soon intruded, and before long it was represented that the tests of significance which could be correctly applied would differ for the two models. (For later developments see a review by Plackett [9]; the discussion on this paper is also worth reading.)

What are the facts? The first and crucial point to recognize is that whether the factor levels are a random selection from some defined set (as might be the case with, say, varieties), or are deliberately chosen by the experimenter, does not affect the logical basis of the formal analysis of variance or the derivation of variance components. Once the selection or choice has been made the levels are known, and the two cases are indistinguishable as far as the actual experiment is concerned. The relevance of the various variance components that can (but need not) be calculated will of course depend on whether the levels can be regarded as approximating to (or are actually) a random selection from some population of levels of interest, but the tests of significance will not be affected.

There is an analogy here with the classical problem of determining the error of a linear regression coefficient, in which, it may be remembered, it was sometimes claimed that allowance had to be made for the fact that the observed $x$ were a sample from some population of $x$, whereas Fisher rightly insisted they could be taken as known.

The difference in definition, though I have not traced it to its source, appears to be in the interaction constants of an $A \times B$ table ($p \times q$ levels, $k$ replicates). If $A$ and $B$ are regarded as random, the cell $(r, s)$ of the table is taken to have a "true" value of

$$\alpha_r + \beta_s + \gamma_{rs}, \tag{5.1}$$

where $\alpha_r$, $\beta_s$ and $\gamma_{rs}$ are members of populations with variances $\sigma_A^2$, $\sigma_B^2$ and $\sigma_{AB}^2$. The marginal mean of the true values for level $r$ of $A$ will then be

$$\alpha_r + \frac{1}{q} \sum_s (\beta_s) + \frac{1}{q} \sum_s (\gamma_{rs}), \tag{5.2}$$

and the mean square for $A$ in the analysis of variance will have expectation

$$\sigma_e^2 + k\sigma_{AB}^2 + kq\sigma_A^2. \tag{5.3}$$

If $B$ is regarded as fixed, the $\gamma$ are redefined so as to have zero marginal means over $B$, that is, to satisfy the conditions

$$\sum_s (\gamma_{rs}) = 0. \tag{5.4}$$

The expectation for the $A$ mean square is then

$$\sigma_e^2 + kq\sigma_A^2. \tag{5.5}$$

This restriction serves no useful purpose. If it is not imposed, the random and fixed models have identical mean square expectations. All that has to be remembered is that the $A$ mean square contains a term in $\sigma_{AB}^2$ as well as $\sigma_A^2$, and similarly for the $B$ mean square. The factorial case then conforms to the convention customarily adopted for a hierarchical classification, where the mean square for a given level contains variance components for that level and all lower levels.

The above differences in expectations have given rise to the belief that the tests of significance differ for the random and fixed models. Denoting the mean squares by $S_A$, $S_B$, $S_{AB}$, $S_E$, we can compare $S_A$ with $S_E$ or $S_{AB}$. The comparison $S_A/S_E$ tests whether for the levels of $B$ in the experiment (whether chosen or obtained by random selection) the effect of $A$ averaged over these levels of $B$ differs from zero. The comparison $S_A/S_{AB}$ tests whether, when the levels of $B$ are a random sample of all possible levels, there is any average effect of $A$ over all possible levels of $B$.

The latter test is clearly not relevant unless the levels of $B$ can be regarded as a random sample of all levels. Even then I would submit that it is pointless, for if there are interaction terms ($\sigma_{AB}^2$ and therefore the $\gamma$ not zero) their averages over all levels of $B$ will in general not be zero, and thus $A$ effects exist, whereas when there are no interactions $S_A/S_E$ is the appropriate test. The appropriate test for interactions is $S_{AB}/\mathrm{S}_E$, or, more conservatively, if $S_A < S_{AB}$, $[S_{AB}, S_A]/S_E$, where [ ] indicates the combined mean square.

Confidence limits for the population average of the $A$ effects can be obtained directly from $S_{AB}$, but here another complication intrudes. If $A$ has more than two levels and there is real variation in the $A$ effects over levels of $B$, there is no reason to expect that $S_{AB}$ will be homogeneous. Thus, if $A$ represents fertilizer levels and $B$ places, the variance of some general measure of response, for example, the linear effect, may be expected to be greater than that of other effect components, for example, the quadratic. Even if $A$ represents varieties the same is likely to be true, for varieties which differ greatly on average may be expected to vary more markedly in their place to place differences. Consequently, for the investigation of place to place differences and similar issues, it is imperative to partition the effects d.f. into single d.f. (usually, but not necessarily, orthogonal) with similar partition of the effects × places d.f. For such investigations to be fruitful fair replication of places is necessary, otherwise the number of effects × places d.f. associated with any particular effect d.f. will be small.

A paper by Harter [7] on the analysis of split plot designs exemplifies the extreme state of confusion that can arise from these differences of definition and from treating replicates as an additional factor $R$ more or less on a par with the treatment factors.

Harter first lists the expectations of the mean squares for $A$ (whole plot factor), $B$ and $R$, with $a$, $b$ and $r$ levels, respectively, and their interactions, in terms of $\sigma_A^2$, $\sigma_B^2$, $\sigma_R^2$, $\sigma_{AB}^2$, $\sigma_{AR}^2$, and so forth, for fixed and random effect models for $A$ and $B$ ("$R$ always regarded as random"). He then gives the list of test ratios shown below, where $\rho$ is the correlation between the subplots in the sense that if $\sigma_w^2$ and $\sigma_s^2$ are the whole and subplot error variances, additional to $\sigma_{AR}^2$ and so forth, $\sigma_w^2 = \{1 - (b - 1)\rho\}\sigma^2$ and $\sigma_s^2 = (1 - \rho)\sigma^2$.

*Test ratios for A.* If $B$ is fixed, use $S_A/S_{AR}$ (exact test). If $B$ is random, use $S_A/S_{AR}$ (assumes $\sigma_{AB}^2 = 0$) or $S_A/S_{AB}$ (assumes $\sigma_{AR}^2 = 0$, $\rho = 0$). Satterthwaite test: $S_A/(S_{AR} + S_{AB} - S_{ABR})$; Cochran test: $(S_A + S_{ABR})/(S_{AR} + S_{AB})$.

*Test ratios for B.* If $A$ is fixed, use $S_B/S_{BR}$ (exact test). If $A$ is random, use $S_B/S_{BR}$ (assumes $\sigma_{AB}^2 = 0$) or $S_B/S_{AB}$ (assumes $\sigma_{BR}^2 = 0$). Satterthwaite test: $S_B/(S_{BR} + S_{AB} - S_{ABR})$; Cochran test: $(S_B + S_{ABR})/(S_{BR} + S_{AB})$.

*Test ratio for A × B.* Use $S_{AB}/S_{ABR}$ (exact test).

If these were indeed the appropriate tests it might, as he states, be a "crucial" question whether or not $S_{BR}$ and $S_{ABR}$ should be pooled to make up subplot error. Far from being crucial, during the many years I have been responsible for the analysis of split plot experiments I have never considered the two components worth separation. The only reason for separation would be to examine, from the results of many experiments, whether there was evidence for variation in response to the subplot factor over replicates; for this purpose $S_{BR}$ might be tested against $S_{ABR}$. This question of differential response has been fairly thoroughly examined in connection with confounded experiments [8], [12], [15] because in such experiments differential response would appear as an apparent interaction between other factors. All the evidence indicates that it is of negligible magnitude in agricultural field trials, even with fertilizer treatments.

Suffice to say that, of the five experiments quoted by Harter, $S_{BR}$ is *less* than $S_{ABR}$ in three, and in none is the difference significant at the 5 per cent level. Yet, following Bozivich, Bancroft, and Hartley [1], he recommends separation in three experiments, in one of which $S_{BR}$ is less than $S_{ABR}$!

Incidentally, it may be noted that Harter's arguments have really little to do with split plots. The same partition of error d.f. can be made in an ordinary $A \times B$ factorial experiment, and the same confused situation would arise if his arguments were valid.

## 6. Nonfactorial response surface designs

Nonfactorial response surface designs, of which rotatable designs are an example, were originally introduced to determine the optimum levels of several factors in industrial processes. They have occasionally been advocated for use in preference to ordinary factorial designs in agricultural field trials whose primary object is to determine optimal levels of fertilizer components. This use seems to me to be very questionable, for the circumstances are very different.

An example of a design that recently came to my notice may bring out the basic objections to rotatable designs. This design was in fact more extreme than a rotatable design, but will serve as an illustration. The design was for a set of trials on nitrogen and phosphate with treatments as in figure 3.

The experimental results were sent to Rothamsted because analysis on a desk calculator was too onerous. It was onerous for us too, as the required formulae had to be worked out, and some special programming was required. This is a very real disadvantage of these designs.

In correspondence on this design we made the following comments. It would have been much better to use the conventional 3 × 3 design, starting each fertilizer at zero level with increments of 30 lb/acre. The information on what happens at the corners of the ordinary conventional 3 × 3 design is of great importance, and in the present design the curvatures are ill determined because of nonorthogonality. The point is illustrated by table I of variances in three replicates of (I) a 3 × 3 factorial design with factors 5 or 35 or 65 lb N per acre, and 0 or 30 or 60 lb $P_2O_5$ per acre, and (II) the design adopted (table I).

TABLE I

COMPARISON OF VARIANCES

| | I | II |
|---|---|---|
| Linear | 1/72 | 1/36 |
| Linear × Linear | 1/192 | 1/12 |
| Quadratic | 1/96 | 5/192 |

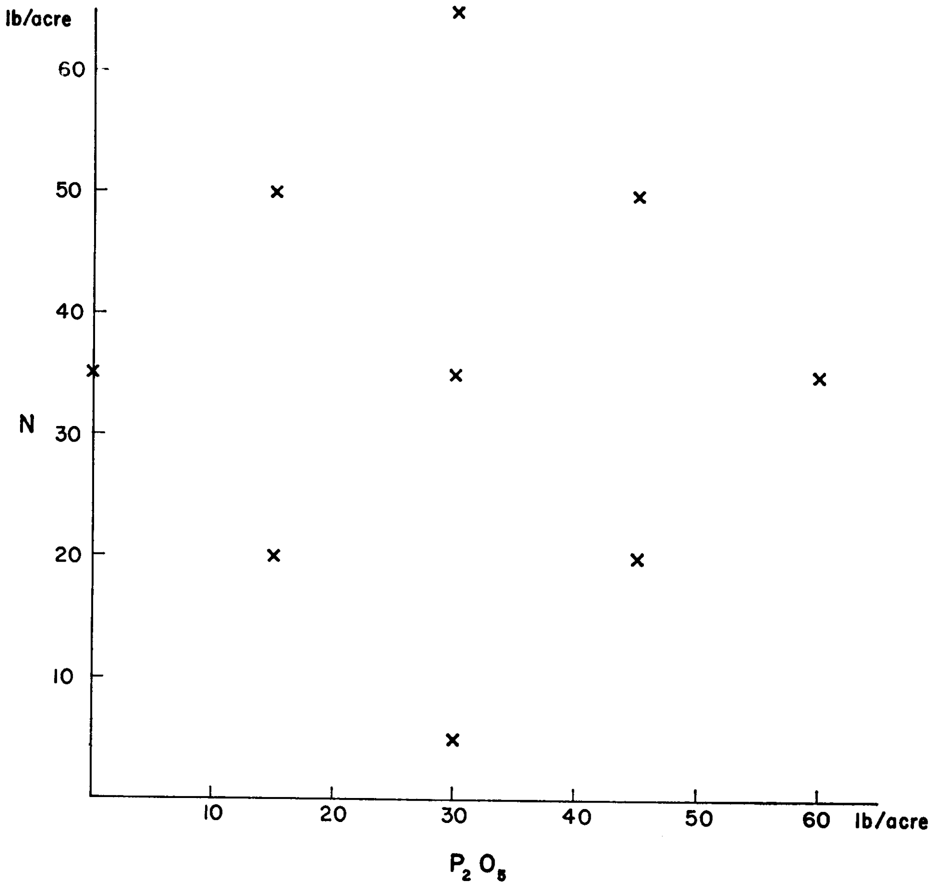


FIGURE 3

Response surface design for N and P.

Of course, Box's designs have advantages over factorial designs. They enable a given number of factors to be tested on fewer units (plots) if necessary, they are particularly suitable for sequential experimentation (indeed they were first introduced for this purpose) and in certain circumstances they are less sensitive to departures from the assumed form of response surface. These points may well be relevant in industrial experimentation, particularly when errors are small, so that loss of efficiency becomes less important, but none is a major consideration in present day fertilizer experiments.

The much lower accuracy of the design is, of course, mainly due to the fact that there are fewer points at the extremes of the permissible ranges of N and P. But for fertilizer components, and I suspect for many factors in other types of experiment, there are no grounds for reducing the range of one factor at the extremes of the others. Exploration of a rectangular area of the response surface, rather than a circular area, is more appropriate. If it were certain that the factors interacted positively, as often happens with fertilizer components, there might be a case for omitting the 0, 2 and 2, 0 combinations in a 3 × 3 design, but negligible interactions are quite common, and in some sets of experiments the existence of strong negative interactions has been established.

The lowest level of a factor need not, of course, be zero. When, for example, it is known from previous work that good responses to N are to be expected, levels of $N_1$, $N_2$, $N_3$ might be adopted for the factorial scheme. It is then useful to include some additional plots (for example, one in each block of a confounded 3 × 3 × 3 factorial) with, for example, no N and the intermediate levels of P and K. This indeed is now common practice, and our computer program for the analysis of 3 × 3 × 3 factorials provides for such additional treatments, and also for summarizing groups of experiments.

One further general point. A single experiment on fertilizers, as indeed is true in much other experimental work, is not expected to give final answers on all points. We do not ask that the optimal dressings should be determinable from one experiment. To do so would demand an impracticably large experiment, and would in any case be relevant only to the chosen field in the current season. All that we expect is reasonably accurate information on the general responses to the separate factors. Information on the curvatures of the response curves and interactions between components, which is required for the calculation of optimal dressings, is gradually accumulated as the experimental program proceeds. This is another fundamental difference between the industrial and agricultural situation.

## 7. Computers and experimental design

Electronic computers are radically altering the computational problems associated with the analysis of experiments, and this has some influence on design. When only desk calculators were available, it was imperative that the arithmetical computations should be kept simple. Now we can, potentially, face much more extensive numerical work if this results in compensating advantages.

I say "potentially" advisedly, for it is unfortunately true that as yet very few computers are programmed to provide full analysis of the more complex types of experiment that are in current use. It may be objected that the need for this is not great, as such experiments were in the past satisfactorily analyzed on desk calculators, and this can continue. This is only partially true, as anyone who has contact with extensive experimental programs knows. Backlogs of unanalyzed results build up, and multivariate techniques involving covariance analysis are seldom used; these are frequently required for experiments in which many variates are observed. Missing observations also cause much trouble.

We have shown at Rothamsted that powerful general programs for the analysis of wide groups of designs can be written. We have, for example, recently written a general factorial program which handles designs with partial (balanced or unbalanced) or complete confounding (including $3^p \times 2^q$ confounded designs), split plots (including successive splits), and even, although this was not planned, fractional replication. Covariance, missing values, and preliminary processing of the data are included, and the results are presented in a form acceptable to those familiar with desk calculators.

I consider that an urgent task facing statistical departments in universities and research institutes is to provide programs of this type on the computers to which they have access. This need not be an onerous task if a cooperative effort is made, and programs are written in a common language such as Fortran or Algol. It would indeed be a great help to practical experimenters if there were a set of standard statistical programs common to all computers of the requisite size. But such standard programs must be good ones. Many of the programs at present available are insufficiently general and unsatisfactory statistically.

To return to the question of how computers will influence design, it is clear that designs requiring the inversion of matrices can now be faced. This is already leading to the development of designs on mixtures (simplex designs); I believe much further interesting work lies ahead of us here. But I would emphasize that in general additional computational complications should not be introduced unless they can be shown to have compensating advantages, and that in many circumstances computationally simple designs, because of their balanced properties, are in fact the most efficient.

Further work for which I believe computers will be particularly useful is the combined analysis of sets of experiments, and the analysis of the accumulated results of long term experiments. The techniques required for these tasks require much further development to which I hope we at Rothamsted will make a useful contribution.

## REFERENCES

[1] H. Bozivich, T. A. Bancroft, and H. O. Hartley, "Power of analysis of variance test procedures for certain incompletely specified models. I," *Ann. Math. Statist.*, Vol. 27 (1956), pp. 1017–1043.

[2] C. Eisenhart, "The assumptions underlying the analysis of variance," *Biometrics*, Vol. 3 (1947), pp. 1–21.

[3] R. A. Fisher, "The arrangement of field experiments," *J. Ministry Agric.*, Vol. 33 (1926), pp. 503–513.

[4] ———, *Statistical Methods for Research Workers*, Edinburgh, Oliver and Boyd, 1925–1958.

[5] ———, *The Design of Experiments*, Edinburgh, Oliver and Boyd, 1935–1960.

[6] P. M. Grundy and M. J. R. Healy, "Restricted randomization and quasi-Latin squares," *J. Roy. Statist. Soc. Ser. B*, Vol. 12 (1950), pp. 286–291.

[7] H. L. Harter, "On the analysis of split-plot experiments," *Biometrics*, Vol. 17 (1961), pp. 144–149.

[8] O. Kempthorne, "A note on differential responses in blocks," *J. Agric. Sci. Camb.*, Vol. 37 (1947), pp. 245–248.

[9] R. L. Plackett, "Models in the analysis of variance," *J. Roy. Statist. Soc. Ser. B*, Vol. 22 (1960), pp. 195–217.

[10] O. TEDIN, "The influence of systematic plot arrangement upon the estimate of error in field experiments," *J. Agric. Sci. Camb.*, Vol. 21 (1931), pp. 191–208.

[11] B. L. WELCH, "On the $z$-test in randomized blocks and Latin squares," *Biometrika*, Vol. 29 (1937), pp. 21–52.

[12] F. YATES, "Complex experiments," *J. Roy. Statist. Soc.*, Suppl. 2 (1935), pp. 181–247.

[13] ———, "A further note on the arrangement of variety trials: quasi-Latin squares," *Ann. Eugen.*, Vol. 7 (1937), pp. 319–332.

[14] ———, "Sir Ronald Fisher and the design of experiments," *Biometrics*, Vol. 20 (1964), pp. 307–321.

[15] F. YATES, S. LIPTON, P. SINHA, and K. P. DAS GUPTA, "An exploratory analysis of a large set of 3 × 3 × 3 fertiliser trials in India," *Emp. J. Exp. Agric.*, Vol. 27 (1959), pp. 263-275.