THE AUTOMATIC DESIGN AND ANALYSIS OF BIOLOGICAL EXPERIMENTS

by

P.J. Claringbold

Department of Veterinary Physiology, University of Sydney

1. INTRODUCTION

In contrast with many of the other problems discussed at this conference the amount of information resulting from a biological experiment is small. Commonly between 50 and 200 observations are made in an experiment. Although there is a growing demand for statistical methods dealing with the multivariate case, i.e. the case of the vector of observations, in most biological experiments only one variable is measured. In addition the magnitude of biological variation and other sources of experimental error is such that refined methods of measurement are quite pointless. A standard deviation of between 20 and 50% of the mean value is common so that accuracy of more than two decimal places is unnecessary. Finally in fields of qualitative or quantal responses the amount of information per observation may be as small as one binary digit.

It is difficult, if not impossible, to generalise on the cost of biological experiments. Animals are usually available in such numbers that experiments of a tactical nature are possible under controlled conditions[1]. Thus the sequential approach[2], of great utility when the cost of making a single trial is high, is not required. The sequential approach has, however, been used in planning of long term agricultural trials and in screening of pharmacological substances[3,4]. Since in tactical experiments the experimenter is free to choose the values of the independent variables and their combinations to be studied full advantage may be taken of the class of experimental designs with orthogonal design matrix. Thus one experiment can yield a great deal of information about the independent variables. For this reason the use of these experiments is becoming more common in biological research.

Until the development of electronic computors computing time was the chief obstacle to the widespread adoption of the designs. It is possible to carry out in a few hours an experiment the complete manual analysis of which may take some weeks. The undesirable nature of such a state of affairs is well recognised, and may now be avoided since automatic computors are programmed.

In order to typify the computational problems encountered in this type of work an experiment is outlined. Suppose 12 experimental variables are varied at either of 2 levels. It is not necessary to complete a full replicate, namely 4 096 = 2^{12} observations, one at each possible treatment combination since the experimenter may assume that interactions between more than two factors are negligible. In this case the experiment is carried out using a few as 256 observations, in 1/16,th replicate. If in addition the block-treatment interactions are assumed negligible the experiment may be performed in four stages, each comprising 64 observations[5]. The problems are:

1. Determination of the admissible treatment combinations.

2. Determination of a randomised permutation governing the correspondence of experimental objects to treatment combinations.

3. Analysis of the results, involving at first $\left[1 + 12 + \binom{12}{2} + 3 \right] = 82$ independent scalar products of the observation matrix and a matrix of comparisons, and then various other matrix operations.

2. OUTLINE OF STATISTICAL METHODS AND NOTATION

Factor levels are specified by integers:-

$$c_i \quad \text{where } i = 1 \text{ through } N. \tag{1}$$

An ordered set of N integers gives a basic description of the experimental design.

Treatment combinations are represented by ordered sets of integers:-

$$(s_i) = (s_1, s_2, \ldots s_N) \tag{2}$$

where s_i is a residue mod c_i.

The treatment combinations in the 2 x 3 experiment are thus denoted,

(00) (01) (02) (10) (11) (12).

An alternative isomorphic notation used by Fisher[6], is not suitable for automatic computing, i.e.:-

(I) b b^2 a ab ab^2.

Treatment effects are similarly represented by ordered sets of integers mod r_i where:

$$r_i \leqq c_i \tag{3}$$

Equation (3) simply states that we may neglect some of the possible treatment effects.

K-matrices or comparison matrices are defined for each experimental factor with a row currency r_i and a column currency c_i, :

$$K_i = \left[k_{t_i}^{s_i} \right] \tag{4}$$

and the elements satisfy the relationship.

$$K_i^T K = \text{diagonal matrix or unit matrix } (1) \tag{5}$$

The rows of comparisons are thus orthogonal, the first row usually being composed of the identity, namely all ones.

Direct product. This matrix operation is of prime importance in the analysis of balanced experimental data. The generalised K-matrix applicable to the data is obtained by the direct product of the matrices defined for each factor. The operation has a simple meaning in the complete replicate of a factorial experiment since there is a one-to-one correspondence between the rows of the product matrix and the effects and interactions of the experiment. In other designs the direct product is only carried out over admissible integer sets.

The direct product is defined:

$$\underline{\underline{K}} = \prod_{\underline{i}=1}^{\underline{N}} \underline{\underline{K}}_{\underline{i}} \qquad (6)$$

where an element of $\underline{\underline{K}}$, $\underline{k}^{(\underline{s}_{\underline{i}})}_{(\underline{t}_{\underline{i}})}$ is given by:

$$\underline{k}^{(\underline{s}_{\underline{i}})}_{(\underline{t}_{\underline{i}})} = \underline{k}^{\underline{s}_1 \underline{s}_2 \cdots \underline{s}_{\underline{N}}}_{\underline{t}_1 \underline{t}_2 \cdots \underline{t}_{\underline{N}}} = \prod_{i=1}^{\underline{N}} \underline{k}^{\underline{s}_{\underline{i}}}_{\underline{t}_{\underline{i}}} . \qquad (7)$$

<u>Admissible treatment combinations</u> are defined in fractional factorial experiments, Latin square and other designs by linear restrictions with integral coefficients on the values of the integer sets. Thus the $\underline{v}$ equations:-

$$\sum_{\underline{i}=1}^{\underline{N}} \alpha^{\underline{k}}_{\underline{i}} . \underline{s}_{\underline{i}} = \underline{Y}^{\underline{k}} \pmod{\underline{p}}. \qquad (8)$$

where $\underline{k}$ = 1 through $\underline{v}$,

$\alpha^{\underline{k}}_{\underline{i}}$ are residues mod p,

$\underline{p}$ is a prime,

define a $\underline{p}^{-\underline{v}}$ replicate of the $\underline{p}^{\underline{N}}$ experiment. Owing to the properties of the Albelian groups fractional replication is only of use in the prime power case and experiments reducible to these by appropriate definition of pseudofactors. Block differences may also be introduced as pseudofactors and associated with high order interactions of the factors, thus:

$$\sum_{\underline{i}=1}^{\underline{N}} \beta^{\underline{m}}_{\underline{i}} . \underline{s}_{\underline{i}} = \mu^{\underline{m}} \pmod{\underline{p}} \qquad (9)$$

where $\underline{m}$ = 1 through $\underline{w}$,

$\beta^{\underline{m}}_{\underline{i}}$ are integers mod $\underline{p}$,

defines the block allocation of the $\underline{p}^{-\underline{v}}$,th replicate into $\underline{p}^{\underline{w}}$ blocks, each block corresponding to one of the $\underline{p}^{\underline{w}}$ possible right hand sides of equation (9). The blocks are usually numbered using residues mod $\underline{p}^{\underline{w}}$.

The coefficients applicable to the problem given in the Introduction of this paper are:

$\alpha^1_{\underline{i}}$ = 1 1 1 1 , 1 0 0 0 , 0 0 0 0

$\alpha^2_{\underline{i}}$ = 1 1 0 0 , 0 1 1 1 , 0 0 0 0

$\underline{\alpha}_{\underline{i}}^{3}$ = 1 1 0 0 , 0 0 0 0 , 1 1 1 0

$\underline{\alpha}_{\underline{i}}^{4}$ = 1 0 1 0 , 0 1 0 0 , 1 0 0 1,

$\underline{\beta}_{\underline{i}}^{1}$ = 1 1 0 0 , 0 0 0 0 , 0 0 0 1

$\underline{\beta}_{\underline{i}}^{2}$ = 0 0 0 1 , 0 1 0 1 , 0 0 0 0.

Admissible treatment effect integer sets: are governed by the assumptions made by the experimenter, i.e. by the associations, if any, between the various factors, and interactions of factors. In the Latin square type experiment main effects are the only effects estimable, while with fractional factorial experiments interactions of more than two factors are usually confounded. As a result treatment effect integer sets are usually considered in an unnatural order, if the numbers they determine as a mixed radix integer is regarded as a natural order. Since it is necessary in automatic computations to be able to generate admissible integer sets some scheme must be evolved. The order is exemplified for the 2^3 case:

000 / 001 010 100 / 011 101 110 / 111 /,

where the bars indicate the end of various interaction classes (-1,0,1and2). The mixed radix case is similar in that the ones are replaced in all possible combinations by other nonzero residues mod $\underline{r}_{\underline{i}}$. For example the 2 x 3 x 2 case:

000 / 001 $\underline{010\ \ 020}$ 100 / $\underline{011\ \ 021}$ 101 $\underline{110\ \ 120}$ / $\underline{111\ \ 121}$ /,

where the underlining indicates related effects.

The binary patterns are thus required in order of decreasing number of zeros and may be so generated if we let:-

$\underline{i}$ = digital position of the first zero in a pattern reading from the left, or = 0 if no such zero exists,

$\underline{j}$ = the digital position of any subsequent one, or = 0 if there is no subsequent one. For example:

	1	2	3	$\underline{i}$	$\underline{j}$	$\underline{d}$
$\underline{s}_0$	0	0	0	1	0	-1
$\underline{s}_1$	0	0	1	1	3	0
$\underline{s}_2$	0	1	0	1	2	0
$\underline{s}_3$	1	0	0	2	0	0
$\underline{s}_4$	0	1	1	1	2	1
$\underline{s}_5$	1	0	1	2	3	1
$\underline{s}_6$	1	1	0	3	0	1
$\underline{s}_7$	1	1	1	0	0	2.

Then $s_{q+1} \pmod{2^N} = 2^i - 1$, when $j = 0$,

$$= (2^i - 1)\, 2^{N-j+1} + [\, s_q \pmod{2^{N-j}} \,]$$

when $j \neq 0$. (10)

The problem is formally identical to listing the vertices of a (N-1) dimensional simplex in order of increasing dimension (d). The procedure may be generalised to describe more complex sequences of integer sets required in more complex analyses, for example:

1	2	3	/4	5	/6	d	
0	0	0	0	0	0	-1	
0	0	0	0	0	1		
0	0	0	0	1	0		
0	0	0	1	0	0	0	6 elements
0	0	1	0	0	0		
0	1	0	0	0	0		
1	0	0	0	0	0		
0	0	0	1	1	0		
0	0	1	0	1	0		
0	1	0	0	1	0		
1	0	0	0	1	0	1	10=15 - 5 elements.
0	0	1	1	0	0		
0	1	0	1	0	0		
1	0	0	1	0	0		
0	1	1	0	0	0		
1	0	1	0	0	0		
1	1	0	0	0	0		
1	1	1	0	0	0	2	1=20 - 19 elements.

The factors are grouped, the bars between the heading numbers indicating the grouping of factors, and a heirachy of parameters used to specify the maximum interaction level within each group. In the example the identity, the 6 main effects, the 10 first order interactions between all but the 6,th factor and the second order interaction between factors 1,2 and 3 are generated. The problem is illustrated by the marked simplexes of the 5-dimensional complex, Fig. 1.

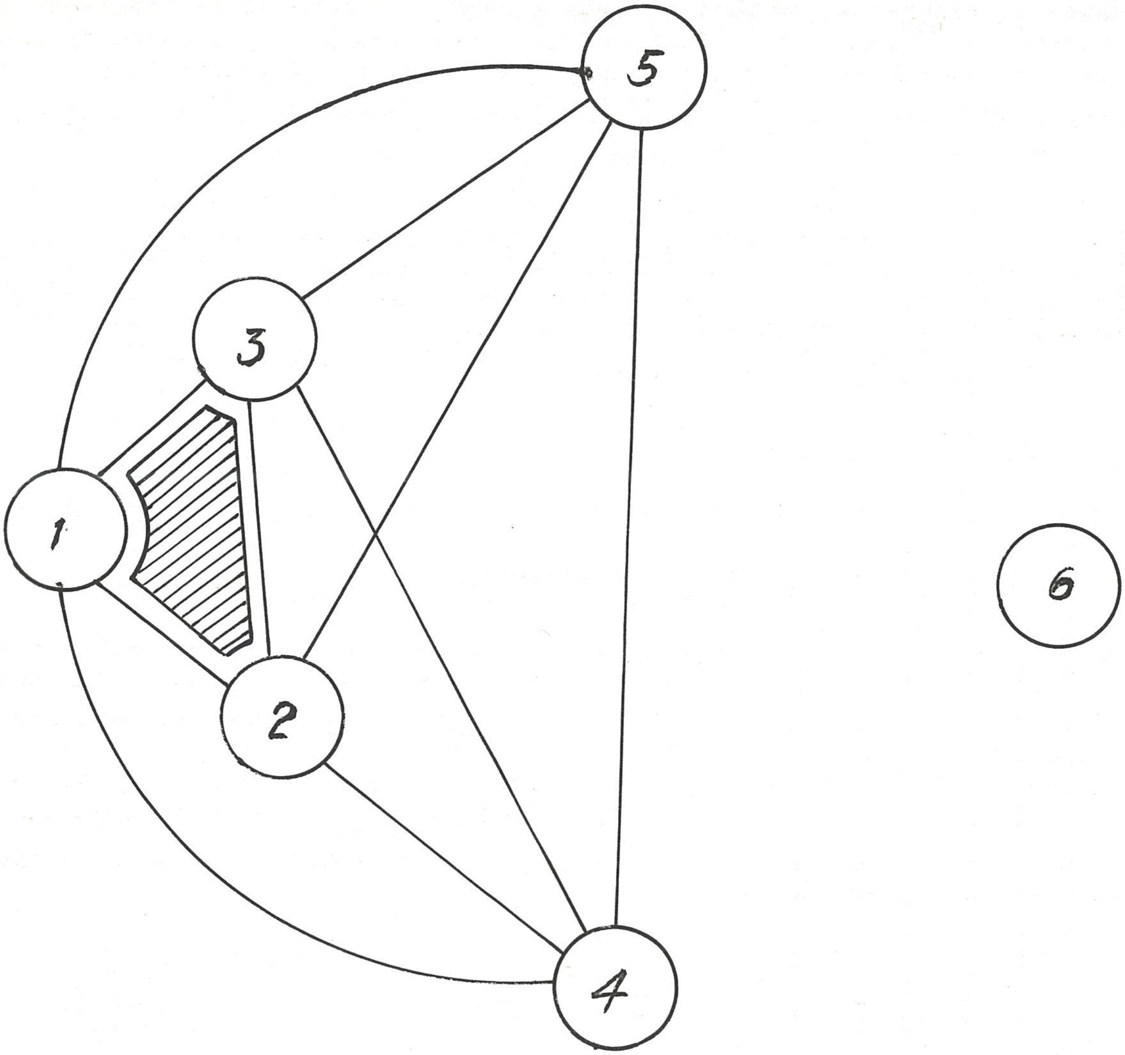

FIG. 1.

Statistical analysis. The statistical methods employed in relating the response variates to the independent variables are those of least squares using the linear model,

$$\underline{\underline{Y}} = \underline{\underline{K}}^{T}\,\underline{\underline{B}} + \underline{\underline{E}} \qquad (11)$$

where $\underline{\underline{Y}}$ is the $\underline{n}$ by $\underline{q}$ matrix of observations,

$\underline{\underline{B}}$ is the $\underline{q}$ by $\underline{g}$ matrix of parameters,

$\underline{\underline{K}}^{T}$ is the $\underline{n}$ by $\underline{g}$ matrix of comparisons,

$\underline{\underline{E}}$ is a matrix of errors.

Since in general a group of observations may be made under identical conditions at each point of the experimental design it is assumed, without loss of generality, that the matrix Y is a matrix of mean values. The matrix of observations takes one of four different forms depending on the nature of the variables measured:

1. A vector if only one discriminant variable is measured,

2. A vector (the discriminant variable) as a column together with a matrix of values of concomitant variables, i.e. variables measured for the prime purpose of increasing the precision of the discriminant variable,

3. A matrix of discriminant variables,

4. A matrix of discriminant together with a matrix of concomitant variables.

In the simple case (1) an analysis of variance may be calculated:

$$(n-g)\ \sigma^2 = (n-g) E^T E = Y^T Y - \sum_{i=1}^{g} B_{(i)}\ K^T_{(i)}\ K^T_{(i)}\ B_{(i)} \qquad (12)$$

where, owing to the orthogonality of the rows of K^T, the sum of squares due to regression is partitioned into g independent squares. The subscript in bracket indicates the i,th row or column of the respective matrices or vectors.

The initial treatment in the remaining cases is identical. Suppose the matrices Y, B and E are partitioned conformably into discriminant and concomitant variables, then corresponding with equation 12 we have the analysis of covariance:

$$(n-g) \begin{bmatrix} E^T E & E^T F \\ & F^T F \end{bmatrix} = \begin{bmatrix} Y^T Y & Y^T Z \\ & Z^T Z \end{bmatrix} - \begin{bmatrix} B^T K K^T B & B^T K K^T \Gamma \\ & \Gamma^T K K^T \Gamma \end{bmatrix} \qquad (13)$$

where the lower blank areas of the matrices are filled by symmetry. The effect of the concomitant variables, cases 2 and 4, on the discriminant variable(s) may be calculated:-

$$R = (F^T F)^I F^T E \qquad (14)$$

and the sum of squares and products of the discriminant variables may then be adjusted for concomitant variation by the rectangular transformation,

$$\begin{bmatrix} I, & -R \end{bmatrix} \begin{bmatrix} B^T K K^T B & B^T K K^T \Gamma \\ & \Gamma^T K K^T \Gamma \end{bmatrix} \begin{bmatrix} I \\ -R \end{bmatrix} \qquad (15)$$

$$= \begin{bmatrix} B^{*T} K K^T B^{*} & -- \\ -- & -- \end{bmatrix}$$

Case 2 has now been reduced to case 1 and 4 to 3.

In the case of multiple discriminant variables (3) analysis may be continued by determination of canonical variables, which are linear combinations of the initial variables rendering the error covariance matrix diagonal under certain restrictions and requirements, usually solutions of the matrix equation:

$$\underline{\underline{S}}_j \, [\, \underline{\underline{A}} - \underline{\underline{O}}_j \, \underline{\underline{E}}^{*T} \underline{\underline{E}}^{*} \,] = 0, \qquad (16)$$

Where $\underline{\underline{A}}$ is a sum of matrices of the form of equation (15).

3. PROGRAMMES AVAILABLE

The programmes discussed below were developed for use with SILLIAC and as a result the approach used is somewhat governed by the facilities and limitations of this computor. At present the prime limitation is the lack of a backing store, long programmes must therefore be read sequentially from paper tape. When magnetic tape becomes available little reorganisation will be required, however, to take full advantage of the new system.

1. Experimental design

Programmes for the listing of admissible treatment combinations, i.e. computing equations (8) and (9) are very simple. As a result additional facilities may be programmed to help in setting up experiments. One point may be mentioned, about the numerical solution of these equations. Digit by digit multiplication and summation is very wasteful of machine time, and may be replaced by <u>whole word</u> logical operations (or multiple word logical operations in the ternary and quinary cases). Thus in the binary case equation 8 becomes a logical product followed by sideways addition of the product i.e. count (mod 2) the number of ones in the resultant pattern.

Within experimental blocks a random correspondence is required between the experimental objects and the admissible treatment combinations for valid tests of significance. If the objects are serially numbered a randomised permutation is required. These are exceedingly tedious to generate by hand from random decimal digits! Since a pseudo-random sequence of integers may be generated by the midsquare method a permutation may quickly be listed and output with the admissible combinations.

Other facilities are programmed also for convenience in preparing data tapes.

2. Data transformation

Many statistical problems may be reduced to the general linear hypothesis (10) by appropriate non-linear transformations of the observed variables. The transformations may be employed either to introduce an additive metric or to render the error covariance matrix independent of location. The transformations commonly used are powers, fractional powers and various logarithmic and exponential functions.

A programme is available which enables transformation of the observed vectors to a new vector, the elements of which are defined functions of the elements of the observed vector. All transformations in common statistical usage are available and simply referable to by code number. Provision is made for omission and rearrangement of input variables on the output tape. Various editorial facilities are also supplied, so that no hand punching of tapes is required after initial punching of the untransformed vectors and the corresponding treatment combinations as tags.

3. Statistical analysis programmes

A number of programmes are available in the section and these enable solution of the four classes of problem discussed on page 10 of this paper.

4. DISCUSSION

In connection with the development of a general programme for the analysis of orthogonal experiment Hartley[7] has suggested an alternative approach to the statistical analysis using the principle of k-way tables. Thus the observation matrix, regarded as a set of vectors by columns, is summed over various combinations of the treatment combination integer sets. Since the number of entries in such tables far exceeds the number of parameters required to describe the data to an equivalent degree of complexity, owing to redundancy of marginal totals, maximal reduction of initial data is not achieved in the first stage of calculation. In the example cited in the introduction, $1 \times 1 + 12 \times 2 + 66 \times 4 + 1 \times 4 = 293$ entries are required in 0, 1 and 2-way tables for complete analysis. These tables may be reconstructed from the 82 entries, one of which corresponds with each row of the K-matrix. Adoption of the K-matrix approach thus results (in this case) in a reduction in intermediate quantities requiring storage by a factor of about 3. Such a reduction is of great importance with a computor such as SILLIAC with only 1024 words available for programme plus the quantities.

Framing of the analysis of covariance in terms of the K-matrix confers additional saving of storage space for intermediate quantities owing to the fact that each row of this matrix is associated with a single degree of freedom in the analysis. As a result the mean squares and products corresponding to a single degree of freedom contrast may be stored as $\underline{q}$ entries instead of $\frac{1}{2}\underline{q}(\underline{q}+1)$ entries. The squares and products may be generated by the direct product of the vector of $\underline{q}$ entries with itself. In all the analysis programmes the matrix product $\underline{\underline{K}}\,\underline{\underline{Y}}$ is calculated, the the elements of $\underline{\underline{K}}$ being generated as required by multiplication of elements listed in the memory. As a result tagged observation groups may be read into the memory in an arbitrary order, the contribution of each group to the sum of square and product matrix and to the $\underline{q}\mathrm{x}\underline{g}$ list of quantities $(\underline{\underline{K}}\,\underline{\underline{Y}})$, is determined before input of the next group of observations. Calculation of the $\underline{g}$ direct products of the sets of $\underline{q}$ contributions followed by normalisation (if $\underline{\underline{K}}\,\underline{\underline{K}}^T \neq \underline{\underline{1}}$) completes this stage of the analysis. If the square and product matrices for each parametric effect require linear transformation for any reason we need only transform the sets of contributions and form the direct product of the new values.

During the development of these programmes for SILLIAC the type of user has been constantly borne in mind. While the statistical methods described above are being increasingly used in biological research, biologists are usually not specialists in numerical analysis as well. A large number of special purpose programmes are clearly undesirable in this situation since an experimenter would not get accumstomed to using a particular programme. A large number of programmes each suitable to a particular experimental design may be feasible if a complete statistical analysis service is supplied together with the computor and some of these have in fact been developed in this situation, e.g. as at the Rothemstead Agricultural Research Station. The additional programming there hardly seems fruitful, however, when it is possible to have a general programme covering many special cases for much the same amount of programming time as a special case.

No.	Author	Title
1.	Fisher, R.A. (1956)	The design of experiments. 6th ed., Oliver and Boyd, London.
2.	Wald, A. (1947)	Sequential analysis. Wiley, New York.
3.	Davies, O.L. (1957)	The design of screening tests in the pharmaceutical industry. Proc. International Statistical Institute, paper 38, Stockholm.
4.	Finney, D.J. (1957)	Statistical problems of plant selection. Proc. International Statistical Institute, paper 69, Stockholm.
5.	Kempthorne, O. (1952)	The design of experiments. Wiley, New York.
6.	Fisher, R.A. (1942)	The theory of confounding in factorial experiments in relation to the theory of groups. Annals of Eugenics, 11, 341-353.
7.	Hartley, H.O. (1956)	Programming analysis of variance for general purpose computors. Biometrics 12, 110-122.

DISCUSSION

Dr. T. Pearcey, CSIRO Sydney

I am rather struck with the possibilities of automatic design in general. Have you given any thought to a more strategic approach?

Dr. P.J. Claringbold in reply

We haven't given a lot of thought to it really. We are more interested in the shape of the response surface and whether set factors are having any effect. We are not interested in minimising some input functions or maximising some product function. For example in experiments of a chemical nature where each observation in general is fairly expensive to make and where the experiments are time consuming, I think a more refined approach is worthwhile. In our case we know we want to find out. We set up the experiment and get answers. If further work is required the initial strategy is altered slightly.

Dr. T. Pearcey, CSIRO Sydney

Presumably the strategy consideration becomes more and more important as the size of the experiment goes up.

Dr. P.J. Claringbold in reply

Yes. Incidentally our approach was based on principle devised by Fox and his co-workers in England.

Mr. T. Newstead, P.M.G. Melbourne

How do you gather and code your data?

Dr. P.J. Claringbold in reply

There are only a couple of hundred observations or observation vectors. As we haven't got a punch handy to our work we write the data down by hand and get someone to punch it out. It would be nice to avoid the writing down stage but it seems hardly worthwhile to go the expense of installing expensive equipment when each experiment is so cheap.