Issues in applying statistical design of experiments, with special reference to toxicology of mixtures

Issues in applying statistical design of experiments, with special reference to toxicology of mixtures

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr. M. Rem, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op dinsdag 17 oktober 2000 om 16.00 uur

door

Eric Dick Schoen

geboren te Zaandam

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. P. van der Laan

en

prof.dr. V.J. Feron

Copromotor:

dr. J.B. Dijkstra

Contents

Chapter 1 General introduction

A scientific experiment may be defined as a planned process of obtaining quantitative information from one or several objects of investigation after applying some controlled treatment. The quantitative information will be referred to as the response to stress the fact that it may respond to changes in treatment. Indeed the treatment may be varied within the experiment to study a possible influence on the responses. Quite often, the treatment can be specified through the settings of several controllable factors. These represent aspects of the experimental conditions that are hold to be potentially influential to the responses.

Following common terminology (see, e.g., Box *et al.*, 1978), a try-out of one specific treatment will be called a run. A statistical design of an experiment is a programme of experimental runs, composed in such a way that systematic effects caused by the controllable factors can be separated from random effects, caused by factors not under the experimenter's control.

Statistical design of experiments is the subject of considerable theoretical development; see for one of many possible examples the combinatorial approach of Dey and Mukerjee (1999). The present Ph.D. thesis, however, is focused on practical applications of statistical designs to scientific research. Special reference is made to the application of existing designs for industrial experimentation to toxicology of mixtures. An example of the application of a statistical design to this field of science is given in Table 1. The experiment bears on the following hypothetical investigation. A toxicologist wants to know whether five mycotoxins, designated A, B, C, D, and E, operate independently on the viability of cells when applied in a mixture. As a measure of cell viability, he selects the ability of the cells to produce some metabolite. The concentration of the formed product can be measured by removing the cells from the surrounding liquid medium, and adding to the liquid a chemical substance that forms a coloured product with the metabolite. Thus, the response is the amount of colour measured in the liquid.

The investigation covers a total of 32 different mixtures of the mycotoxins. The mycotoxin are controllable factors with a concentration from a predetermined set of two called the levels of the factor. The 32 mixtures are just all the combinations of five factors and two levels per factor. As such, it is called a full factorial design. An important alternative is the fractional factorial design, in which only a part of the full factorial design is carried out.

The complete set of 32 mixtures enables the calculation of 31 effects, viz., five main effects, ten two-factor interactions, ten three-factor interactions, five four-factor interactions, and one five-factor interaction. Their calculation is as follows. The main effect of a factor is calculated as the difference in the means taken at the respective levels of the factor. For example, the main effect of mycotoxin B is calculated as the mean of all observations made at the high concentration of this mycotoxin, minus the mean of the remaining observations. The other main effects are calculated analogously.

Two-factor interactions are differences in main effects. The interaction between the mycotoxins A and B, for example, is calculated as the difference in the main effect of mycotoxin A at the high concentration of mycotoxin B with the main effect of mycotoxin A at the low concentration of mycotoxin B, divided by two. The divisor causes all interactions

to have the same standard error as the main effects. The remaining kinds of interactions are calculated in a way analogous to the calculation of the two-factor interactions. A full treatment is given in Box *et al*. (1978).

One element of a statistical design is the list of factor combinations to be performed. Another element bears on the experimental conduct, succinctly indicated in Table 1 by the columns 'plate' and 'well'. 'Plates' are tissue culture plates with small containers ('wells') to hold cell tissue in a liquid medium. The hypothetical example uses eight plates of eight wells each. Separate treatments can be applied to each of the wells. The statistical design

Table 1. Example of an experimental design

mycotoxin[a]					position			
A	B	C	D	E	plate	well	plate	well
1.1	0.11	54	2.4	1.75	3	8	8	5
1.1	0.11	54	2.4	3.25	2	2	4	7
1.1	0.11	54	3.4	1.75	2	5	4	1
1.1	0.11	54	3.4	3.25	3	2	8	6
1.1	0.11	82	2.4	1.75	5	7	6	8
1.1	0.11	82	2.4	3.25	1	3	7	2
1.1	0.11	82	3.4	1.75	1	5	7	3
1.1	0.11	82	3.4	3.25	5	6	6	5
1.1	0.23	54	2.4	1.75	1	2	7	1
1.1	0.23	54	2.4	3.25	5	2	6	7
1.1	0.23	54	3.4	1.75	5	4	6	6
1.1	0.23	54	3.4	3.25	1	4	7	8
1.1	0.23	82	2.4	1.75	2	1	4	6
1.1	0.23	82	2.4	3.25	3	5	8	3
1.1	0.23	82	3.4	1.75	3	6	8	2
1.1	0.23	82	3.4	3.25	2	6	4	3
2.3	0.11	54	2.4	1.75	1	6	7	6
2.3	0.11	54	2.4	3.25	5	8	6	3
2.3	0.11	54	3.4	1.75	5	3	6	4
2.3	0.11	54	3.4	3.25	1	1	7	4
2.3	0.11	82	2.4	1.75	2	8	4	5
2.3	0.11	82	2.4	3.25	3	4	8	8
2.3	0.11	82	3.4	1.75	3	3	8	4
2.3	0.11	82	3.4	3.25	2	4	4	4
2.3	0.23	54	2.4	1.75	3	1	8	7
2.3	0.23	54	2.4	3.25	2	7	4	2
2.3	0.23	54	3.4	1.75	2	3	4	8
2.3	0.23	54	3.4	3.25	3	7	8	1
2.3	0.23	82	2.4	1.75	5	5	6	2
2.3	0.23	82	2.4	3.25	1	7	7	7
2.3	0.23	82	3.4	1.75	1	8	7	5
2.3	0.23	82	3.4	3.25	5	1	6	1

[a]Concentrations given in nM (C) or μM (A, B, D, and E).

specifies the distribution of the treatments over the plates and the wells. This particular distribution is carried out in such a way that three out of the 31 effects to be calculated are affected by random differences between the plates. The remaining 28 effects are not affected by between-plate differences, but only by random differences between the wells of a plate. As various aspects of the experimental conduct are carried out on plates as a whole, it is reasonable to assume that the random between-plate differences will be more substantial than the random within-plate-between-well differences.

The statistical design of Table 1 has four sets of two plates with identical sets of treatments. Thus, differences between the mean response of the plates in each of these sets quantify the random between-plate variation. There are also eight differences between identically treated wells of a set of two plates. The mean difference estimates the between-plate variation; with this difference subtracted from the set of eight differences, we are able to quantify the random within-plate differences.

The above example was constructed to illustrate the following four issues in applying statistical design of experiments. Firstly, the research problem should be translated in statistical terms. Thus, applying a statistical design always involves considerations of the field of research as well of statistical considerations. In the example, the problem of establishing the mode of operation of five mycotoxins was translated in the statistical terminology of main effects and interactions. Absence of interactions indicates an independent mode of action.

A second issue is the construction of a set of treatments. This entails a decision on using factorial designs. If such designs are to be used, then one has to decide on the levels of the factors and the combinations of factor levels to be used in the experiment. Usually, the levels - or at least the range of these - are chosen on the basis of earlier experience. The choice of combinations of the levels is typically a statistical issue, although there may be practical constraints (see standard textbooks on applied statistics, e.g., Box *et al*., 1978; Montgomery, 1997). In the example, there is to be a two-level experiment, possibly in view of cost considerations. The factorial character permits a sensitive detection of active interactions. All combinations of the factors' levels are tried out, permitting the unambiguous estimation of all main effects and all interactions. All the treatments are tried twice. This feature permits the estimation of random error.

Thirdly, there is the issue of allocating the treatments to the experimental units. This issue is intimately connected to the experimental conduct. In the example, the plates had 8 wells each. A possible option is to allocate the treatments at random to the total number of available wells. This option would avoid any systematic connection between treatments and the position of the wells on some plate. However, it was anticipated that wells of one and the same plate would be more homogeneous than wells across plates. So it is sensible to exploit the homogeneity within the plates. The example had three effects of lesser importance confounded with plate differences, while the remaining effects are not. For a non-technical discussion on randomisation, see Cox (1958), and Lorenzen and Anderson (1993).

Finally, as will be shown more fully in Chapter 2, the statistical analysis of the results can also be viewed as a design issue. This is because the separation of random and systematic differences highly depends on the exact handling of the experimental material and the available treatments. To illustrate this, consider as a possible alternative of the toxicological experiment the use of a half-fraction of a full experiment, in four replicates. The half-

fraction used most frequently permits estimation of all five main effects and all ten two-factor interactions, if it can be assumed that there are no interactions of an order higher than one. As the alternative experiment has as many wells as the experiment from Table 1, but a smaller number of systematic effects, the design has more room for the estimation of random variation. The point to be made here is that these considerations may well be used to make the choice between the different designs. If there is special interest in the random between-plate and within-plate variation, then the alternative design could be attractive. If it were suspected that there are complex interactions, then one would prefer the original design.

The hypothetical experiment and its discussion reflect the close connection between statistical design of experiments, practical considerations from the field of application and considerations of the experimental conduct. The connection features in all of the eight studies collected in the present PhD thesis. These studies reflect the range of the author's activities as a statistical consultant, as they cover the development of new statistical methods as well as the application of existing statistical designs for industrial experimentation in toxicology of mixtures. The new methods give practical solutions to the problems of dealing with several nested sources of uncontrolled variation in two-level experiments, using two-level designs to handle three-level and four-level factors, and analysing effects from unreplicated experiments by formal procedures. The existing designs enable toxicologists to detect toxicological interactions in mixtures of chemical agents under minimisation of the number of animals, the time of experimentation, and the experimental costs.

The remaining part of this thesis is organised as follows. Chapter 2 is on the construction of fractional designs. The subsequent statistical analysis of these designs and other unreplicated designs is the subject of Chapter 3. When statistical designs and their subsequent analysis are fully worked out, they become part of a standard repertoire of methods. Chapter 4 features applications of designs from the standard repertoire of industrial experimentation to toxicology of mixtures. All the studies in the Chapters 2-4 feature practical examples of statistical design of experiments. In the Chapter 5, the findings of these examples are evaluated in the light of the availability of the particular statistical techniques that were used to reach the conclusions. References to current literature are collected in a separate section following the chapters. Finally, there are 3 Appendices in Dutch containing a summary of this thesis, acknowledgements, and a brief overview of the scientific career of the author.

The remaining part of this introduction gives an overview of the Chapters 2-4 and the publication status of the studies constituting these chapters.

Construction of fractional designs

To reduce the size of a screening experiment, factors are often studied at two levels only. However, there may be categorical factors that can attain more than two levels. If the number of possible levels is only slightly more than two and if there are just a few of this type of factor, then it is sensible to include all the levels in the experiment in combination with two-level factors. It then becomes a problem how to design a fraction of a full design

and how to analyse the results. Study 2.1, titled *Design and analysis of a fractional $4^1 3^1 2^5$ split-plot experiment*, covers the particular case of an experiment with a four-level and a three-level factor, in combination with some two-level factors. The four-level factor is constructed by combining the levels of two two-level factors (Wu, 1989). The four combinations of the two-level factors jointly define the levels of the four-level factor. This is illustrated in Table 2.

One possible way of constructing a three-level factor from two level factors is to construct first a four-level factor and to combine two levels of this factor to a single level of the three-level factor. The three-level factor thus has one level with twice as many observations as each of the remaining levels. Study 2.1 uses a different mode of construction. Here, one factor from a fractional two-level design is chosen to cover two of the levels of the future three-level factor. This design is then augmented with some half fraction of the parental design.

In Study 2.1, the proposed mode of analysis of the experimental results is the half-normal plotting of effects (Daniel, 1959). To construct a half-normal plot, the absolute values of the effects are ordered and plotted against some statistical function of the order numbers. The function is constructed in such a way that effects with true values equalling zero form a straight line through the origin of the plot. Effects clearly deviating from the line are marked as active.

In the study, it is proposed more in particular that the half-normal plotting is to be carried out for the three subsets of the results formed by taking pairs of levels of the three-level factor. Therefore, it makes sense to evaluate the overall design properties by looking both at the properties of these pairs and at those of the design at the separate levels of the three-level factor. The evaluation should consider all possible designs resulting from the choices for the two-level factors to construct the four-level factor, and for the two-level factor to construct the three-level factor. The particular evaluation criterion proposed in Study 2.1 is the aberration criterion of Fries and Hunter (1980) as modified by Wu and Zhang (1993) for mixed four-and-two level designs. This is a systematic way of looking at the series of effects on which no information can be obtained due to the fractionation of the design. Using this criterion, it was possible to discard one of three $4^1 3^1 2^5$ designs as inferior to the others. However, more work is needed to formulate a single criterion for all subsets of the results formed by taking pairs of levels of the three-level factor, in stead of applying the criterion for each subset separately.

The hypothetical experiment from the start of this introductory chapter had two sources of random variation, viz., those between plates and between wells within a plate. In addition there are effects affected by the between plate random variation and effects affected by the other source of random variation. The analysis of this type of experiments is

Table 2. Construction of a four-level factor from two two-level factors[a]

two-level factor I	two-level factor II	four-level factor
1	1	1
2	1	2
1	2	3
2	2	4

[a]Figures are coded factor levels.

standard (see, e.g., Montgomery, 1997).

Unreplicated experiments with two sources of random variation are a bit more difficult to analyse. The set of effects is split-up according to the random variation affecting the effects. If the number of effects in a set is sufficiently large, say seven or larger, then we may try to separate a few really active effects from a majority of inactive ones. Study 2.1 shows that we can adapt fractional designs from a catalogue for a single source of random variation to cases with two ore more sources of random variation in such a way that each set of effects is sufficiently large. This problem is worked out more systematically in Study 2.2, titled *Designing fractional two-level experiments with nested error structures.* One part of the study is concerned with allocating a given set of treatments to the experimental material in such a way that the number of distinct sources of random variation is minimised. This is important because the effects fall apart in separate sets according to these sources. Minimisation of the number of distinct sources increases the mean number of effects in any particular set. It then becomes easier to separate active from inactive effects, if one can assume that only a few of the effects are indeed active.

In stead of adapting the experimental material to a given set of treatments, one could also adapt the set of treatments to a given set of experimental material. Study 2.2 formulates a heuristic step-by-step procedure for this purpose. The procedure indicates how to use existing catalogues of two-level designs to construct the required design. The procedure does not claim optimality properties according to some criterion, but it is believed to result in satisfactory designs in terms of aberration.

Analysis of unreplicated designs

The effects in experiments with a single source of random variation can, after a suitable standardisation, all be evaluated with one and the same standard error. Designs with a nested error structure – such as the ones from Chapter 2 - have the additional complication that the effects fall apart in two or more sets. One of these sets has effects that are affected by just a single source of random variation, while the effects from the other sets are affected by additional sources of variation. Thus, separate standard errors for each of the sets are to be established. In the absence of replication, there are several approaches to construct the standard errors from the sets of effects. One of these is the halfnormal plotting presented in Chapter 2. A clear disadvantage of this approach is the subjectivity as to which effects belong to a straight line through the origin, and which of them do not. For this reason, numerous more formal procedures have been developed. Haaland and O'Connell (1995) unified several of these procedures by their common elements. They presented constants needed to calculate standard errors with two of three recommended procedures, and critical values for tests with one of these, each for experiments with selected numbers of effects in the range from 7 to 31. Study 3.1, titled *Three robust scale estimators to judge unreplicated experiments*, extends the results of Haaland and O'Connell (1995) to constants and critical values for all effect numbers in the range from 7 to 127 and all three estimators. To aid in the practical use of the estimators, it is argued that the choice of the estimators be based on the capability of the design to distinguish main effects from two-factor interactions and to distinguish two-factor interactions from each other.

The results of Study 3.1 were used to analyse the 31 effects from a two-level design with 5 factors in *Assessment of some critical factors in the freezing techniques for the cryopreservation of precision-cut rat liver slices* (Study 3.2). This study consists of two sub-studies. The initial one focuses on detection of overall differences between two techniques, while the second one aims at detecting the critical factors explaining the differences with the aforementioned two-level design with 5 factors. The design is seemingly replicated because of the presence of pairs of identically treated liver slices. However, as these slices are treated simultaneously, the relevant standard error for the effects is the one between the pairs of slices. As the treatments are carried out only once, the statistical evaluation of the effects uses the procedure from Study 3.1. It is shown that residual viability of liver slices after cryopreservation and subsequent culturing is mainly determined by freezing rate and the cryopreservation medium. One particular combination of these factors was identified as the most promising approach to successful freezing of rat liver slices.

Study 3.2 shows that the detection procedure from Study 3.1 is useful in its own right. The effects detected with this procedure are those affecting the mean, or location of the response of interest. However, the procedure may also be used as the first step in the detection of effects of the experimental factors on the uncontrolled variation around the mean. The detection of these so-called dispersion effects is important both in industry and in toxicology of mixtures. For industry, factors having a dispersion effect may be set in such a way as to minimise the dispersion. This may result in products showing a more constant set of properties. For toxicology of mixtures, a dispersion effect of one of the factors may indicate that the experimental material does not respond uniformly to some treatment. This may either be caused by uncontrolled variation in applying the treatment or by biological variation of the experimental units. Either of these modes of action may be exploited further. For example, treatments may be modified to get a more thorough action, while a change in biological variation to a treatment may need to be incorporated in risk assessment.

Dispersion effect detection is addressed in Study 3.3, titled *Choosing appropriate two-level experiments to detect location and dispersion effects*. For unreplicated experiments, we first detect location effects with a method from Study 3.1. These are then subtracted from the value of the response. What remains is used as a measure of uncontrolled variation in the experiment. In the second step of the procedure, a statistical test is applied to detect factors affecting the uncontrolled variation. Recent literature shows a severe sensitivity of the test to unidentified location effects. Using the results from replicated experiments has the obvious advantage that we have a direct access to uncontrolled variation via the replicate variances. Therefore, this type of experiments is also included in the study. In the study, practical conditions are derived for which the sensitivity of the tests in unreplicated experiments to unidentified location effects is less severe. For further evaluation, we present a large-scale simulation study covering two-level experiments with 16 to 64 runs and up to 9 factors. Conditions under which dispersion effects can be reasonably detected in unreplicated screening experiments are too restrictive to be recommended, although the use of an alternative strategy for one of the tests studied may improve its performance. For replicated designs, a single large dispersion effect can be detected with 32 runs, while 64 runs permit detection of a medium and a large effect.

Statistical designs in toxicology of mixtures

Fractional designs, such as the ones studied in the Chapters 2 and 3, are typically applied in industrial settings. The driving force behind the development of these designs is cost minimisation. In toxicological research, these designs are not well known. However, research on the combined effect of mixtures of toxic compounds often has practical limitations as to the number of distinct treatments in an experiment. This is due to the daunting number of possible mixtures of interest, while there are also practical limitations of the experimental equipment due to the complexity of the study. As there is usually a considerable biological variability in the experimental material, it is important to gain by replication what is lost by limiting the number of treatments. Thus, it would be natural to use replicated fractional experiments in mixture toxicology. Accordingly, Study 4.1, called *Statistical designs in combination toxicology: a matter of choice*, explains features of fractional designs useful for mixture toxicology. It demonstrates that, contrary to the impression one may get from published literature, there are often several alternative designs for a given toxicological problem. The choice between these alternatives is a matter of balancing an increase in complexity of experimental conduct against the gain in useful information when a more complex experiment would be favoured.

The practical examples discussed in Study 4.1 include a toxicity study of nine chemicals in rats. Here, the choice between two designs with 9 factors at two levels favoured the practical feasibility of 16 treatment groups of five animals each over the gain in information when 32 treatment groups with less animals each would be chosen. A full account of the 16-treatment design is given in Study 4.2, titled, *Subacute toxicity of a mixture of nine chemicals in rats: detecting interactive effects with a fractionated two-level factorial design.* Indeed, one of the purposes of this study was to evaluate the use of fractional designs to detect interactions between the chemicals in a mixture.

In the aforementioned design, each of the chemicals was either absent in the mixture or present at its 'minimum-observed-adverse-effect level' (MOAEL). The design permits the calculation of 6 combinations of two-factor interactions. Of the 29 response variables of main interest, 16 showed activity of one or more of these combinations. By taking into account the results on main effects and expert knowledge obtained from previous studies, the interactions that were likely to be active could be identified.

The fractional design in Study 4.2 was applied without previous experimental evidence of interactions. An attractive alternative would be to carry out some preliminary experiment to establish the presence of interactions first, and following up with further experiments to clear up which of the possible interactions are active. In the previous *in vivo* study, this type of extensive testing is far beyond financial and experimental possibilities. The use of a relatively simple *in vitro* testing system, however, allows the investigation of many experimental groups. This notion is illustrated explicitly in Study 4.3, *Statistically designed experiments to screen mixtures of* Fusarium *mycotoxins for possible interactions.* Here, a testing strategy is developed to detect interaction between the toxic mycotoxins. In an initial experiment, a mixture of the mycotoxins is applied in various dilutions to cell cultures, while there are also treatments with single mycotoxins in various concentrations. As a measure of the cell proliferation, the extent to which radioactive material is incorporated in the cells is studied. The design used permits the detection of interaction

between the mycotoxins through comparing a predicted activity from the individual mycotoxins with the observed activity of the mixture of mycotoxins. The statistical testing procedure has been specifically developed for this purpose. The second experiment in the strategy uses a central composite design to reveal the interactions that are indeed active, together with curvilinear relations in single factors. In the third experiment in the strategy, the effects detected in the central composite design are tested for final assessment in full factorial designs with two factors each.

The above strategy is illustrated with experimental results for each of the stages. In the first stage, cells treated with the mixture at its highest concentration showed a response that was less than additive. Therefore, some of the mycotoxins must interact. The follow-up revealed four interactions, which were, surprisingly, more than additive. Two of these were investigated in the final step. One interaction was confirmed, the other was not confirmed. It was concluded that the effect of the mixtures could not be predicted solely on the basis of the individual mycotoxins.

Publication status

The eight studies collected in this PhD thesis are slightly altered versions of published papers or manuscripts of papers submitted for publication. The publication status of the respective papers and manuscripts is given below.

2.1 Schoen, E.D., and Wolff, K. (1997) Design and analysis of a fractional $4^1 3^1 2^5$ split-plot experiment. *Journal of Applied Statistics* **24**: 409-419.

2.2 Schoen, E.D. (1999) Designing fractional two-level experiments with nested error structures. *Journal of Applied Statistics* **26**: 495-508.

3.1 Schoen, E.D., and Kaul, E.A.A. (2000) Three robust scale estimators to judge unreplicated experiments. *Journal of Quality Technology* **32**: 276-283.

3.2 Maas, W.J.M., de Graaf, I.A.M., Schoen, E.D., Koster, H.J., van de Sandt, J.J.M., and Groten, J.P. (2000) Assessment of some critical factors in the freezing technique for the cryopreservation of precision-cut rat liver slices. *Cryobiology* **40**: 250-263.

3.3 Schoen, E.D., and Siersma, V.D. (submitted) Choosing appropriate two-level experiments to detect location and dispersion effects.

4.1 Schoen, E.D. (1996) Statistical designs in combination toxicology: a matter of choice. *Food and Chemical Toxicology* **34**: 1059-1065.

4.2 Groten, J.P., Schoen, E.D., van Bladeren, P.J., Kuper, C.F., van Zorge, J.A., and Feron, V.J. (1997) Subacute toxicity of a combination of nine chemicals in rats: detecting interactive effects with a two level factorial design. *Fundamental and Applied Toxicology* **36**: 15-29.

4.3 Tajima, O., Schoen, E.D., Feron, V.J., and Groten , J.P. (submitted) Statistically designed experiments in a tiered approach to screen mixtures of *Fusarium* mycotoxins for possible interactions.

Chapter 2 Construction of fractional designs

2.1 Design and analysis of a fractional $4^1 3^1 2^5$ split-plot experiment

Introduction

Fractional factorial designs are used where a full factorial experiment would take too much effort in time or money. They can be used fruitfully when the number of degrees of freedom for potentially interesting effects is but a fraction of the available degrees of freedom in a complete experiment. For example, a complete 2^8 experiment comprises 255 degrees of freedom. Only 36 of them involve main effects and first-order interactions. Under the assumption that all interactions with order higher than 1 are negligible, we would prefer to use the one-quarter fraction of Box *et al*. (1978, p. 410) rather than a complete replicate. The fraction gives unbiased estimates of the 36 potentially relevant effects together with 27 residual degrees of freedom. We may even prefer to use the one-eighth fraction from Box *et al*. (1978, p. 410) giving unbiased estimates of the main effects, but not of all first-order interactions (there are six pairs and one triplet of aliased interactions; 13 interactions are unbiased; the remaining three degrees of freedom involve higher order interactions).

As regards the practical use of two-level fractional designs, a cursory examination of the literature on applied statistics suffices to demonstrate their popularity. However, examples of fractional designs with all factors at three levels, do not seem to be abundant in the literature. It is not difficult to find reasons for this. In view of the larger number of degrees of freedom involved with interactions, relatively large experiments are required to leave the first-order interactions unconfounded (Daniel, 1976, p. 220). Moreover, when the factors are quantitative, response surface methods (Box and Draper, 1987) seem greatly superior.

For qualitative factors, response methods do not apply. We then have no choice but to use multi-level factorial designs. When such a design involves factors that differ in their number of levels, it is called a 'mixed design'.

Mixed designs are particularly easy to construct if the numbers of factor levels in a design are all powers of two. For example, McLean and Anderson (1984, pp. 66-70) showed that an eight-level factor may be thought to consist of three 'pseudo-factors' that have two levels (for a more general treatment of pseudo-factors, see Monod and Bailey, 1992). The 'main effects' and the 'interactions' of the pseudo-factors jointly describe the main effect of the eight-level factor. Thus, a mixed power-of-two design is effectively reduced to a two-level design. For the special case of $4^1 2^p$ and $4^2 2^p$ designs, optimal implementation of the multi-level factors is discussed by Wu and Zhang (1993). Connor and Young (1961) gave plans of mixed designs in the $2^p 3^q$ series. These are such that all the main effects and first-order interactions can be estimated. The designs are slightly non-orthogonal. Higher-order interactions are assumed to be negligible. Their absence allows estimation of random error.

This study bears on mixed designs that involve two-level factors (2LFs), a three-level factor (3LF), and a four-level factor (4LF). We used such a design to investigate a new method to amplify DNA. In the remaining part of this study, we will discuss the following issues. We start with practical aspects of the investigation. Originally, the experiment was a $4^1 2^6$; the treatments will be discussed next. It was natural to confound the main effects of two of the factors with blocks. We will discuss confounding an additional effect to enable the evaluation of whole-plot effects with half-normal plots. In the experiment, two devices to multiply DNA were to be tested. After preparation of the experimental plan, a third device became available for testing. In view of restrictions in time we copied the runs of one of the original devices with the third one. This is discussed in a further part of the paper. The analysis of the results is given next. Subsequently, we study four ways to design a $4^1 3^1 2^5$ experiment from a 2^8. They are classified according to (1) the original 2LF to become a three-level one and (2) the relation between the runs at each of the levels of the 3LF. The aliasing in the designs is evaluated with the methodology for $4^1 2^p$ designs developed by Wu and Zhang (1993). Some brief remarks on other issues conclude the paper.

Experimental design

Practical aspects

Isolation of DNA from individual organisms yields only very small quantities of material for further study. Therefore, biochemical techniques have been developed to amplify the original DNA. Williams *et al*. (1990) introduced one such technique, called the Random Amplified Polymorphic DNA (RAPD) technique. The experimental procedure is roughly as follows. The DNA from which specific fragments are to be amplified, is placed in a reaction-vial, to which several other ingredients are added. A number of vials are placed together in a heating-and-cooling device called a thermocycler; the device generates temperature cycles during which an enzym (polymerase) amplifies the DNA. After amplification the composition of the DNA is analysed by gel electrophoresis. The yield is recorded through visual comparison of stained gels with a standard of known content. Details can be found in Wolff *et al*. (1993).

Our investigation of the RAPD technique focuses on the effects of several conditions on the yield. The following factors were likely to be critical: the kind of thermocycler; the annealing temperature (lower temperature of the cycle); the enzyme brand; the magnesium concentration; the DNA primer used to start the DNA synthesis; the pH value; and the gelatin concentration. The first two factors affect the temperature cycling process; the remaining five factors affect the ingredients in the vial.

We had as many as 20 primers at our disposal; four of them were actually used. This is a compromise between the wish to investigate more than two primers and the wish prevent too much entangling of effects (see Discussion). If the remaining factors were studied at two levels, then one replicate of the design would still require as many as 256 observations. Therefore, we looked for a fractional $4^1 2^6$ design. Table 3 shows the factors, their levels, and the design with which their effects are investigated. The design will be discussed more fully later in the study.

Four-level factor

For the construction of our $4^1 2^6$ experiment we first chose a convenient 2^8 plan. Then, we combined two of the factors in the design yielding a 4LF (Addelman, 1962; Wu, 1989). In the following, we call these factors pseudo-factors (see Monod and Bailey, 1992).

Our two-level design was the 2^{8-3} design from Box *et al*. (1978). Its defining relation is

$$I = ABCF = ABDG = ADEFH = CDFG = BCDEH = BEFGH = ACEGH.$$

The design has resolution IV, because the shortest word in the relation contains four letters. The four-letter words imply that there are main effects that are aliased with second-order interaction. Also, there are first-order interactions aliased with each other.

All designs of the collection in Box *et al*. (1978) have minimum aberration (Fries and Hunter, 1980): the number of words with minimum length is minimal. We have here three words of four letters; the remaining four words have five letters. For an experiment in which no factor is *a priori* known to have an effect, a minimum aberration design with resolution IV seems a reasonable choice because, given the number of observations, it minimises the seriousness of the aliasing.

Wu and Zhang (1993) extended the concept of minimum aberration to designs with 4LFs and 2LFs. For $4^1 2^p$ designs, they suggested that a single letter be used for the product of the two pseudo-factors. They made a distinction between words containing no pseudo-factors (type 0) and words containing one pseudo-factor (type 1). Type 1 words may be considered as less disturbing than type 0 words. A minimum aberration design of type 0 contains the smallest number of type 0 words of minimum length (an extension to cover equal numbers of type 0 words is straightforward).

The minimum aberration $4^1 2^6$ design given in Wu and Zhang (1993) contains five words

Table 3. Factors, levels and design of the thermocycler experiment[a]

Factor	-1 level	+1 level
A: thermocycler	I[b]	III
B: annealing temperature	35 °C	36 °C
C: enzyme brand	brand 1	brand 2
D: magnesium concentration	1.5 mM	2 mM
E: DNA primer, pseudo-factor 1	primers 3 and 14	primers 4 and 20
F: DNA primer, pseudo-factor 2	primers 3 and 4	primers 14 and 20
G: pH	8.3	7.7
H: gelatin concentration	0.001 %	0.01 %

[a] 2^{8-3} design with F = ABC, G = ABD, and H = ADEF. Block generators: A, B, ACD. [b] Treatments with thermocycler I are duplicated with thermocycler II.

of four letters; one is of type 0, four are of type 1. There are also two type 1 words with five letters. The authors represented the word-length pattern with {(0,0),(1,4),(0,2)}. The pairs give the number of three-, four-, and five-lettered words, respectively. The first figure of a pair relates to the words with no pseudo-factor, and the second number relates to words with a pseudo-factor. Products of pseudo-factors are represented with one letter. We can make our design have exactly the same pattern of word lengths when we choose either 'E' or 'H' as one of the pseudo-factors. The choice gives resolution IV because these letters only occur in the five-lettered words of the above defining relation. The other pseudo-factor should be neither E nor H, otherwise seven words of four letters would result.

We chose E and F as pseudo-factors. The resulting design can be obtained by relabelling the factors in the minimum aberration design of Wu and Zhang (1993). See Table 3 for further assignment of letters to the experimental factors.

Use of half-normal plots; confounding

All the vials of a particular run of the thermocycler have the same level both for the factor thermocycler and for the annealing temperature. Therefore, we have a split-plot experiment with the thermocycler runs as whole plots and vials as sub-plots. This suggests a division of the 31 effects into three between-runs effects (annealing temperature, thermocycler and the interaction between annealing temperature and thermocycler) and 28 within-runs effects.

The statistical evaluation of unreplicated split-plot experiments was exemplified by Box and Jones (1993). They used separate normal plots for the between-runs effects and the within-runs effects; the effects were plotted against the inverse standard-normal score of their cumulative percentage points. Effects with expectation zero are roughly on a straight line through the origin; prominent effects 'fall off' the line. We prefer the original half-normal plots (Daniel, 1959), because these are insensitive to the arbitrariness of sign of the effects (Loh, 1992). We will discuss evaluation of the set of effects with a standard error constructed from the same set in Study 3.1.

Using a (half-)normal plot it is difficult to detect a strong between-runs effect when there is only a choice between three effects . By confounding yet another effect between the runs, we realised a division in seven between-runs effects and 24 within-runs effects. This is because the generalised interactions between thermocycler, annealing temperature, and the third effect are also between-runs effects.

The choice of the third effect is not arbitrary. The effects A, B, and (ABH + CFH + DGH), for example, would yield A x B x ABH = H as one of their generalized interactions. Because we want to measure main effects more precisely than interactions, this is not a very good combination of terms. Given the assignment of A and B to the thermocycler and annealing temperature, respectively, a better choice for the third effect is CD. This results in the confounding of A, B, AB + CF + DG, CD + GF, CG + DF, EH, and a string of aliased second-order interactions, respectively, with runs.

Three-level factor

In our experimental design the thermocycler is a 2LF. Its original levels were I and III. Shortly after the preparation of the experimental plan, a third thermocycler, designated II, became temporarily available for testing. Because there was little time to spare, the intuitive solution to duplicate the runs of device I with the new device was adopted. This measure is evaluated in later in the paper.

The design resulting from jointly considering all three thermocyclers is not orthogonal. Therefore, we evaluated the results of pairs of thermocyclers; we also considered each thermocycler singly.

Analysis of results

The data of the experiment are given in Table 4. In a preliminary analysis of the results from thermocyclers I and III, we fitted one and the same model for untransformed data, square-rooted data, and logarithmically transformed data. For this purpose, the two zeros in the data were replaced with 0.0001. The model contained all effects that were conspicuous in any of the half-normal plots. The square-rooted data were compatible with the usual assumption of homogeneity of variances. (These data will be used in the following.)

Table 4. Results from the thermocycler experiment[a]

D = -1	D = +1	D = -1	D = +1
E = -1	E = -1	E = +1	E = +1
0.4 1.6	1.1 2.0	1.2 2.0	2.0 2.65
0.9	6.0	0.9	2.0
1.2 2.0	0.9 2.6	2.5 2.0	1.0 1.9
4.0	3.0	2.5	4.0
0.7 0.004	0.001 2.0	0.05 0.001	0.4 2.0
0.01	0[b]	0.1	0[b]
0.5 1.5	0.3 1.3	0.2 1.2	2.5 1.9
1.3	0.2	1.1	1.1

[a]Yields of DNA in μg per 30 μl reactionproduct. See Table 3 for experimental design. Columns are in Yates order for factors A, B, and C. Pairs give results for thermocycler I (first) and II (last); single figures bear on thermocycler III. [b]Replaced with 0.0001.

Active effects

Half-normal plots of between-runs effects that relate to the three pairs of thermocyclers are given in Fig. 1. The plot for thermocyclers I+III suggests that the largest effect may be judged real. It appears to be an effect of B (annealing temperature; value: 0.42). The same value is obtained by combining the results of thermocyclers II and III , but, in the relevant plot the effect is masked by the ACD effect (value: -0.35). In the plot for thermocyclers I+II, it is obscured by its aliases (B - CF - DG; value: 0.25). We conclude that B is probably active.

Half-normal plots of within-runs effects that relate to the three pairs of thermocyclers (not shown) suggest various active effects. Those that involve main effect C and interaction BF appear in every set; they are probably active. In addition, each thermocycler had active effects of its own. This implies interactions between thermocycler and the other factors. Therefore, the results of all three thermocyclers cannot conveniently be pooled.

Comparison of error variances

We constructed separate models for the results from the pair (I, III) and for those from (II), respectively. Table 5 shows the analysis of variance. The variance within runs for (I,III) differs from the corresponding variance for (II) (F test; $p = 0.004$). Therefore we did not pool the residual variances. (A graphical check on the results of (I,III) indicated compatibility with the assumption of homogeneity of variances).

An explanation of the difference in within-runs variance could be that only thermocycler II uses reaction vials specifically designed for this device. Thus thermocyclers I and III may not have optimally constructed reaction vials.

Both sets of data showed significant additional variation due to runs. The F-values were 3.03 (I,III) and 28.52 (II).

Table 5. Analysis of variance for experimental data[a]

		I+III		II	
		df	MS	df	MS
between runs	model	1	1.4425	1	0.2534
	error	6	0.3184	2	0.3513
within runs	model	4	1.6341	4	0.5783
	error	20	0.1052	8	0.01232

[a]Square-rooted data from Table 2. Roman numerals denote thermal cyclers. Model terms for I+III: B (between runs); C, AC+BF, BH+EFG, ABH (within runs). Model terms for II: B-CF-DG (between runs); C-BF, D-BG-EH, F-BC, G-BD (within runs).

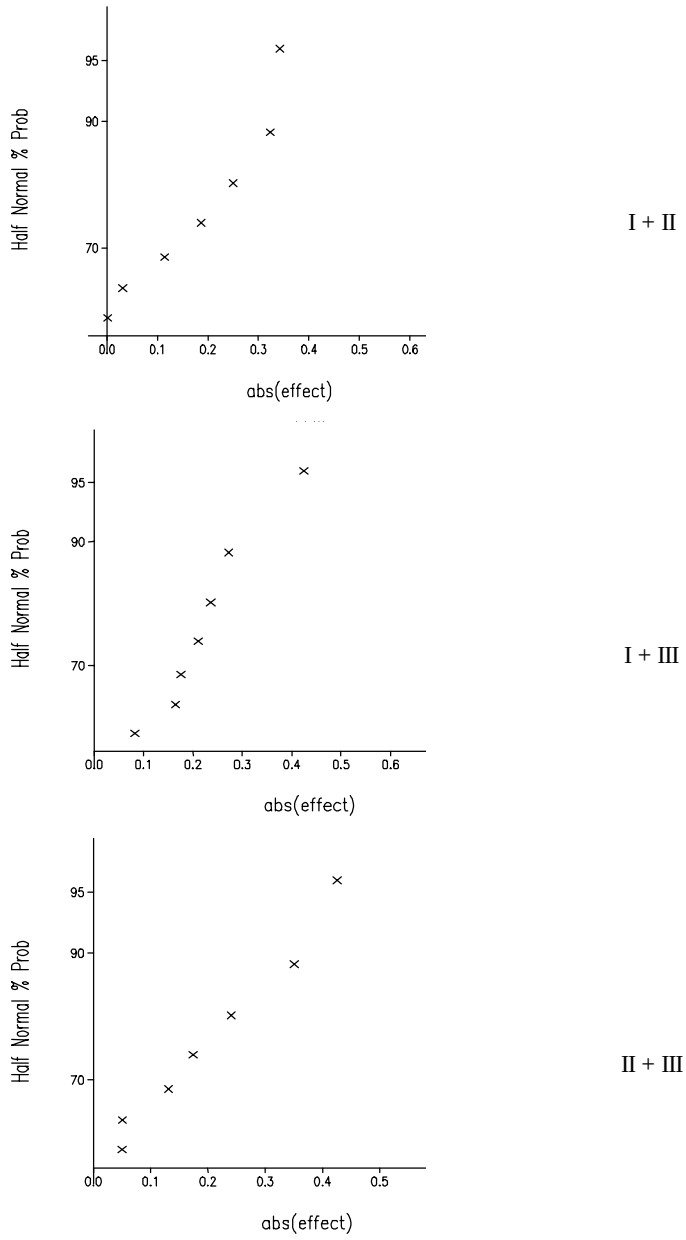


Fig. 1. Half-normal plots of the between-runs effects for pairs of thermocyclers

1

De-aliasing

We may combine the results from (I,III) with those from (II) to solve an ambiguity in the effects. The BF effect for (I,III) is confounded with the AC effect. The sum of the effects is -0.32. From the results obtained with thermocycler II, we obtain the effect of C - BF. Its magnitude is -0.47. On adding this to the previous effect, we obtain an estimate of -0.79 for the effect of C in (II) plus that of AC in (I,III). This value is very close to the value for C obtained with (I,III). We conclude that the value of -0.32 for AC + BF results from BF.

Three-level factor in a minimum aberration design

The generators of the design for single thermocyclers (signs not considered) are BDG, BCF, and DEFH, respectively. The design is clearly of resolution III. The word-length pattern, in the notation of Wu and Zhang (1993), is {(1,2), (0,3), (0,1)}. The minimum aberration $4^1 2^5$ experiment of these authors, however, has pattern {(0,2), (1,4)}. Clearly, our intuitive solution to copy the runs of thermocycler I with thermocycler II does not lead to a minimum aberration design for the single thermocyclers. We now show how a more judicious choice of the pseudo-factors leads to $4^1 2^p$ designs that exhibit better properties with regard to aberration both in singles and in pairs of thermocyclers.

In the defining relation of our original (see above), all letters appear in 4 words. The letters E and H appear only in each of the 4 five-letter words; the remaining letters appear in 2 four-letter words and 2 five-letter words. Thus there are 2 basically different choices for allocation of a factor to a letter, say ‘A’or ‘H’.

Table 6 shows the word-length pattern for singles and pairs of thermocyclers for each of the above choices for the factor thermocycler, when the design would be viewed as a pure two-level design. We also show how the signs of the generators ABDG, ABCF, and ADEFH, used to define the design for individual thermocyclers, affect the word-length pattern. In particular, there is a difference between giving the generators of thermocycler I and II identical signs, opposite to those of thermocycler III, and giving each thermocycler a distinct sign pattern for its generators. The ‘equal’ signs correspond to copying the runs of thermocycler I with II, the ‘unequal’ signs are inspired by the mode of construction of $2^p 3^q$ experiments by Connor and Young (1961).

From Table 6 the choice of H for the factor thermocycler leads to a minimum-aberration two-level design for single thermocyclers irrespective of the sign pattern for the generators. In addition, the choice of unequal signs, gives a minimum aberration two-level design for each of the pairs of thermocyclers. Therefore, the mode of construction of Connor and Young (1961) is also optimal from the two-level minimum aberration point of view.

We now choose pseudo-factors in each of the four designs of Table 6. We note first that the words BDG and BCF and their product CDFG are in the defining relations of single thermocyclers when A is chosen to represent a thermocycler, irrespective of the equality of signs for I and II. We can minimise aberration when we make both these words of type 1. The obvious way of doing this is to choose B for one of the pseudo-factors. To avoid a resolution II design, we must not choose the other pseudo-factor from C, D, F, or G. Either

E or H as the second pseudo-factor gives the design for single thermocyclers minimum aberration.

When considered in pairs, there is a difference between the word pattern for the equal and unequal cases with letter A for a thermocycler. The word-length pattern for the 'equal' case is given in design (1) of Table 7; the 'unequal' case corresponds to design (2) of Table 7.

The choice for H as the thermocycler yields a defining relation that consists of seven four-letter words (Table 6). If the devices are considered individually, then it is immaterial what choice we make for the pseudo-factors. When we consider pairs, we should make different choices for the 'equal' and 'unequal' cases. Looking at the pairs I+II and II+III in the 'equal' case, the letter A appears only in five-letter words. Each pair from the remaining letters appears in one of the four-letter words. Therefore, A and any other letter will keep the design aberration minimal. The choice A + E yields design (1) of Table 7.

For the unequal case with H for the thermocycler, the pair I+II has F, I+III has D, and II+III has B only in five letter words. Obviously at least one of these letters should be chosen as a pseudo-factor. The choice of two of them (B+D) yields design (3) of Table 7, the choice of one of them (B) with another letter (E) yields design (2).

Of the three designs in Table 7, design (3) obviously has less aberration of type 0 than does design (1). The classification of design (2) is less clear. It has only one minimum aberration pair, but the two other pairs exhibit less aberration than the individual pairs of

Table 6. Number of three-, four-, five-, or six-lettered words in four derivations from a basic 2^{8-3} design, split up in singles and pairs of thermocyclers[a]

signs of generators for I and II	thermocycler combination	letter for factor thermocycler							
		A				H			
equal	single	2	3	2		0	7		
	I + II	2	3	2			7		
	I + III		3	4			3	4	
	II + III		3	4			3	4	
unequal	single	2	3	2		0	7		
	I + II	1	2	3	1		3	4	
	I + III	1	2	3	1		3	4	
	II + III		5	0	2		3	4	

[a]Roman numbers denote thermocyclers. Basic design has generators ABDG, ABCF, ADEFH. Derivations are classified according to letter used for factor thermocycler in basic design and signs of the generators for thermocyclers I and II (equal: - - -, - - -; unequal: - + -, - - +; thermocycler III has + + +).

(1) and (3) not having minimum aberration. Therefore, the choice between (2) and (3) remains somewhat arbitrary.

Surprisingly, design (2) of Table 7 can be achieved through the choice of A as well as of H for the thermocycler. We conclude that the aberration properties of all pairs of levels of the 3LF are greatly affected by the choice of the 2LF that represents two levels of the 3LF, while unequal signs for the generators yield the superior designs.

To the author's knowledge, the proposed way of using the aberration criterion for the mixed-four-and-three-and-two level experiment has not appeared elsewhere in the literature. More work is needed, however, to develop a single criterion for the whole of the design. Such a criterion would permit the construction of systematic catalogues of $4^1 3^1 2^p$ designs.

Table 7. Word-length pattern in the designs of Table 6 after assignment of pseudo-factors[a]

	signs of generators for I and II	letter for factor thermocycler	pseudo-factors	thermocycler combination	word-length pattern		
(1)	equal	A	B, E	single	**0,2**	**1,4**	
	equal	H	A, E	I + II	0,2	1,4	
				I + III		**1,4**	**0,2**
				II + III		**1,4**	**0,2**
(2)	unequal	A	B, E	single	**0,2**	**1,4**	
	unequal	H	B, E	I + II	0,1	0,3	1,2
				I + III	0,1	0,3	1,2
				II + III		**1,4**	**0,2**
(3)	unequal	H	B, D	single	**0,2**	**1,4**	
				I + II	0,2	1,0	0,4
				I + III		**1,4**	**0,2**
				II + III		**1,4**	**0,2**

[a]Number of three-, four-, and five-lettered words with 0 (first figure of a pair) or 1 (second figure) pseudo-factors. Product of pseudo-factors is treated as one letter. Patterns of minimum aberration are printed in bold. For further explanation, see Table 6.

Discussion

This section offers some concluding remarks on (1) the capability of the present experiment to cope with machines behaving differently, (2) the number of primers chosen for the experiment, (3) the analysis of any resolution IV 3^q2^r experiment, and (4) the use of extra effects for confounding in split-plot experiments.

The purpose of the present experiment was to earmark from a predetermined list the aspects of the DNA amplification that affect the yield. One of the items on the list was the thermocycler. Because it is common for pieces of equipment to behave differently, interactions of the active factors with the thermocycler were to be expected. The experimental design afforded detection of such interactions because, in every pair of thermocyclers, interactions with the thermocycler were unconfounded with main effects.

It is normal to investigate the active factors in more detail in a second experiment. This was indeed done for thermocycler II, although the device actually used in the new experiment was somewhat different. Wolff *et al*. (1993) give more details.

We had as many as 20 different primers at our disposal. The use of eight or 16 of them would involve three or four pseudo-factors with all of their interactions. Even if there were no interactions between the primers and the remaining factors, we would have to use two-level designs with resolution VI (eight primers) or VII (16 primers) to keep an acceptable aliasing pattern. The experiment would then become far too costly to perform. The choice of four primers yields a design of acceptable resolution at an acceptable number of observations.

This paper demonstrates the analysis of a 3^12^p design with half-normal plots for each individual level of the 3LF. When more than one 3LF is needed, the analysis may become more complicated. Separate half-normal plots may be made of each combination of the 3LFs, but the number of effects may become very small. Also, we may extend the mode of analysis presented in this paper, such as by making quadruplets of the nine treatments generated by two 3LFs. This mode of action can be adopted when the design is not fractional regarding the 3LFs.

Finally, we wish to remark on the confounding in the present study. We deliberately confounded a third effect to be able to evaluate between-runs effects. It is difficult to see how the probable effect of B could have been established if there were only three between-runs effects in the results from (I,III). For fractional factorial designs with several error strata, we therefore recommend the inclusion of at least three independent effects in each stratum. We might even have done better if the experiment had had 16 runs of size two. The mean variance of the 31 effects would have been slightly more than with eight runs of size four, but the between-runs effects would be more numerous and more precise. Of course, the disadvantage is the doubling of the time needed for the experiment. Clearly, there is a trade-off between practical and statistical considerations.

2.2 Designing fractional two-level experiments with nested error structures

Introduction

Industrial research often requires screening of a large number of experimental factors to detect which of them is influential and which of them is not influential. For such an investigation, a fractional two-level experiment (Box *et al.*, 1978) may be well suited. If interactions are negligible, it enables the study of up to 2^n - 1 main effects in 2^n runs. Also, such an experiment is the starting point for the construction of mixed two-and-four level experiments (Wu, 1989).

For practical reasons, the experiment may be conducted in such a way that a nested error structure ensues. We will detail three instances. First, the process under consideration may dictate a split-plot structure of the experiment. In cheese making experiments, for example, some treatment factors are applied to whole milk supplies, while others are applied to curds productions. There are several curds productions (corresponding to subplots) made from a single milk supply (corresponding to whole plots). A reasonable model for an unreplicated experiment of this kind is

$$u_{ij} = \beta_0 + x_1^T \beta_1 + e_i + x_2^T \beta_2 + e_{j(i)} \qquad (1)$$

In this formula, u_{ij} is the response of subplot j in whole plot i, $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are column vectors with the settings of the explanatory variables varied between the whole plots and between subplots within the whole plots, respectively, β_0 is a general mean, β_1 and β_2 are column vectors with the true effects of the explanatory variables, and e_i and $e_{j(i)}$ are random contributions of whole plots and subplots, respectively.

A nested error structure may also be the result of a convenient grouping of experimental runs. Indeed such a measure increases the precision of the estimated effects (Cox, 1958). If the grouping factor is not related to the treatment factors, then the groups are usually called blocks. As an example, several runs in a chemical experiment could be performed simultaneously on a single day. Thus, the factor 'day' defines the blocks of the experiment.

If the blocks can be considered as a random sample from some population, or the blocks are allocated at random to sets of treatments, then it is reasonable to model the contributions of the blocking factor as random effects. The resulting formula is identical to (1), with a change in terminology: 'blocks' are used for 'whole plots' and 'units' or 'runs' are used in stead of 'subplots'.

As a final instance of experiments with nested error structures, blocking could also be applied in the presence of typical whole plot factors. The preceding chemical experiment, for example, could include a factor that requires the treatment of experimental units with a chemical. The treatment is applied by exposing several units in a jar. The whole plots defined by the jars may or may not coincide with the blocks defined by the days. Thus we may need an extension of model (1) with an additional error term to model the jars, an additional vector of fixed effects varied between the jars within a day, and a vector of the corresponding true effects.

The literature on designing fractional two-level experiments with nested error structures concentrates almost exclusively on blocking. The usual approach here is to look for a set of independent effects called 'blocking generators' to be confounded (see, for example, Montgomery, 1997). This approach presupposes a decision on the number of blocks. However, in many instances, various blocking options for a given set of treatments are practically feasible, and statistical guidelines are needed to further the choice.

For split-plot fractional two-level designs, the catalogue of Addelman (1964) is the only reference known to the present author. It does not give explicit construction principles, however. For cases not covered in the catalogue, one could try to use some main effects in a given design as 'blocking' generators. This seemingly straightforward solution may result in an incompatibility with the split-plot structure (see later). We conclude that explicit principles are needed to deal with such an incompatibility.

Finally, there is a lack of literature on combining blocking with whole plots and subplots in a fractional factorial design. In conjunction with the required guidelines for blocking options and the explicit principles to deal with split-plot features, this motivates the formulation of a general strategy to handle fractional two-level designs with nested errors. This study proposes such a strategy. Two practical examples illustrate its key elements. The first example is a 32-run chemical experiment with 16 two-level factors. Eight reaction vessels may be used simultaneously, so various blocking options are possible. There are also whole plot treatment factors that may introduce additional grouping of the experimental material.

The second example is a 128 run cheese-making experiment that has 11 two-level factors. There is a split-split-plot structure with sets of individual cheeses nested within curds productions, and curds productions nested within milk supplies. Addelman (1964) does not cover this case, while a standard 2^{11-4} design (Box *et al.*, 1978) is shown to be incompatible with the split-plot features of the experiment.

The organisation of this study is as follows. The examples are introduced in the next section. Subsequently, a review of terminology for fractional designs with nested errors is given. The examples are then used to develop the strategy. The designs for both examples are completed in another section. The study is concluded by a discussion.

Examples

A chemical experiment

The Department of Chemical Engineering of Delft University tried to find from a list of potentially influential factors those that really affect the yield of a catalyst synthesised on gauze. The two-level factors being used specify changes in the experimental procedure

that consist of the following stages (see Table 8). First, gauzes are prepared by cutting strips from a roll and treating them according to one of various protocols, as classified by two factors. Secondly, the ingredients for the chemical synthesis are mixed. Nine factors characterise the mixture. The four combinations of the two-level factors 'Si source 1' and 'Si source 2' are to define four different Si sources (see Wu 1989). Similarly, four Al sources are to be defined by the factors 'Al source 1' and 'Al source 2'.

In the third stage of the experimental procedure, the mixture is shaken thoroughly and there is a short resting time for the mixture prior to further handling; there are two factors bearing on the treatment of the mixture. Subsequently, the mixture is put in a reaction vessel called an 'autoclave', and heated for several hours either rotating or standing. This entails another factor. An experimental run is concluded with cooling down of the

Table 8. Experimental procedure for chemical example[a]

stage of procedure	factor	levels
I gauze preparation	etching (H)	yes, no
	pretreatment (J)	yes, no
II mixing components	Si source 1 (D)	1, 2
	Si source 2 (E)	1, 2
	Al source 1 (G)	1, 2
	Al source 2 (P)	1, 2
	template (K)	1, 2
	Si:Al ratio (L)	high, low
	template:Al ratio (M)	high, low
	total salt (N)	high, low
	water:Al ratio (Q)	high, low
III treatment of mixture	shaking time (A)	long, short
	aging (B)	long, short
IV synthesis	rotating (C)	yes, no
V end of synthesis	cooling (O)	forced, unforced
	time after synthesis (F)	long, short

[a] Letters between brackets are codes for the factor names.

autoclave according to the setting of a factor, waiting before opening of the autoclave according to another factor, the rinsing of the gauze and the determination of its weight gain.

In this experiment, there were a total of 16 two-level factors to be screened. It was decided that a total of 32 runs should suffice. With regards to blocking, it is possible to use eight identical autoclaves simultaneously. Thus, 32 blocks of one synthesis (no blocking at all), 16 blocks of two, eight blocks of four, and four blocks of eight are the options. Those taking the least amount of time, i.e. 4 x 8 and 8 x 4 will be considered more fully. The problem is to choose between these options.

A cheese-making experiment

A cheese-making factory constantly receives milk supplies from farmers. On entering the factory, the milk is stored in tanks. At an appropriate time, the contents of a tank are emptied in smaller tanks in which the curds is made. A single storage tank can fill about 10 production tanks. Finally, the contents of each curds production tank are transported to cheese presses to produce many individual cheeses. If a cheese-making experiment entails factors for each of the production-stages, then it is clearly of the split-split plot type.

In a two-level experiment for a European cheese-making factory (represented in Fig. 2), two factors worked on whole milk supplies. Two curds production tanks were available to investigate four factors relating to curds making. It seemed sensible to include the factor 'production tank' as well. Therefore, five factors were to be varied between the curds productions of the same milk supply. Finally, four more factors could be applied on sets of cheeses of one and the same curds production. Two of these factors were to be combined to a four-level factor.

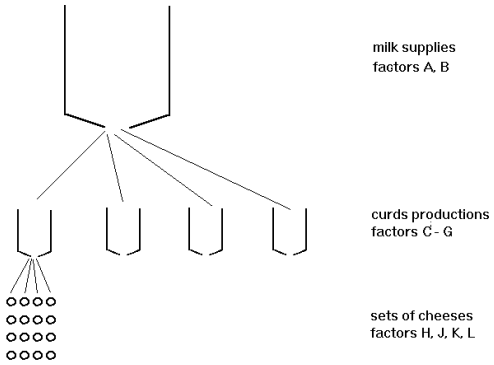


Fig. 2. Structure of the cheese-making experiment

The purpose of the experiment was to detect which of the factors affected the quality characteristics of the cheeses and to obtain an appreciation of the amount of interaction between the factors. As a result of various restrictions, the experiment had to be performed with at most eight milk supplies, using at most four curds productions per milk supply, and using at most four sets of identically treated cheeses per curds production. The problem here was to design an experiment with these maximum numbers which maximises the number of distinguishable two-factor interactions.

Review of terminology

A two-level fractional experiment of f factors in $N = 2^n$ units, can be designed by coding the levels of all the factors with -1 and +1, respectively, writing down a full design for n factors and equating each of the remaining $p = f - n$ factors to either +1 or -1 times an effect involving some or all of the former n factors. The design is said to have p 'treatment generators' defined as the signed product of one of the p factors with the effect to which it was equated. A generator equals +1 for all experimental units. The design has a 'defining relation', defined as the set of 2^p 'words' consisting of the identity, all generators, and all products involving 2 up to p generators. The 'resolution' of the design is the length of the shortest word in the defining relation.

A two-level design with f factors in 2^n units is of 'maximum resolution' if the shortest word in the defining relation is of the largest possible length (Box and Hunter, 1961). Fries and Hunter (1980) extend this idea by defining designs with 'minimum aberration' (MA). Roughly, a design is MA if the number of smallest words in the defining relation is minimal. See the original paper for a more precise definition.

A two-level design with a nested error structure of $B = 2^b$ blocks, or whole-plots, of 2^k units, or subplots, has a set of b independent effects called 'blocking generators' to define the blocks. The set of the blocking generators and all products involving 2 up to b blocking generators jointly comprise the between-blocks (or whole-plot) effects.

An 'error stratum' of an experiment with a nested error structure is defined as a set of all the contrasts with one and the same precision. Thus, a split-plot experiment has a whole plot and a subplot stratum, while a blocked experiment has a 'between blocks' and a 'within blocks' stratum. The lower stratum is defined to be the error stratum associated with the least number of variance components, the *upper stratum* has all components of the nested error structure.

Strategy

This section presents a practical strategy to decide on an experimental design for two-level fractional experiments with nested error structures. Two features of these experiments require special attention. Firstly, their contrasts are divided over several error strata. As a consequence of the division, the usual evaluation with halfnormal plots (Daniel,1959) has to be applied for each error stratum separately (Box and Jones, 1992; Schoen and Wolff, 1997; see also Study 2.1). As an example of separate halfnormal plotting, Fig. 3 presents

the plots of the chemical experiment for the natural logarithm of weight gain. The contrasts are divided over two strata corresponding to contrasts between days, and between gauzes within days. Four large contrasts are marked. More details on the design are given later.

A second feature peculiar to two-level fractional experiments with nested error structures, is the possible incompatibility of a given set of treatments with the error structure. The cheese making experiment, for example, is not compatible with all possible 2^{11-4} designs. The predetermined error structure requires 32 curds productions for the investigation of 7 factors. The 2^{7-2} subdesign needed here has at most a resolution of IV, which is in contrast with the maximum resolution of a 2^{11-4} design, being V (Box *et al*., 1978).

The practical strategy proposed here aims to create error strata with sufficient numbers of contrasts to separate active effects from inactive effects. In addition, the strategy details the construction of treatment generators, given the constraints of a predetermined error structure. These issues are worked out more fully in the following.

Error strata with sufficient numbers of contrasts

Active contrasts from unreplicated experiments may be separated from inactive contrasts by halfnormal plotting. However, such a separation is difficult to achieve when there are few contrasts in a plot. This raises a question as to a minimum number of contrasts for a halfnormal plot to work well. It is difficult to give well-founded rules here, but published examples rarely show plots with less than seven contrasts. Thus, we propose the following.

Guideline 1. A halfnormal plot should have a minimum number of seven contrasts to enable a sensible evaluation.

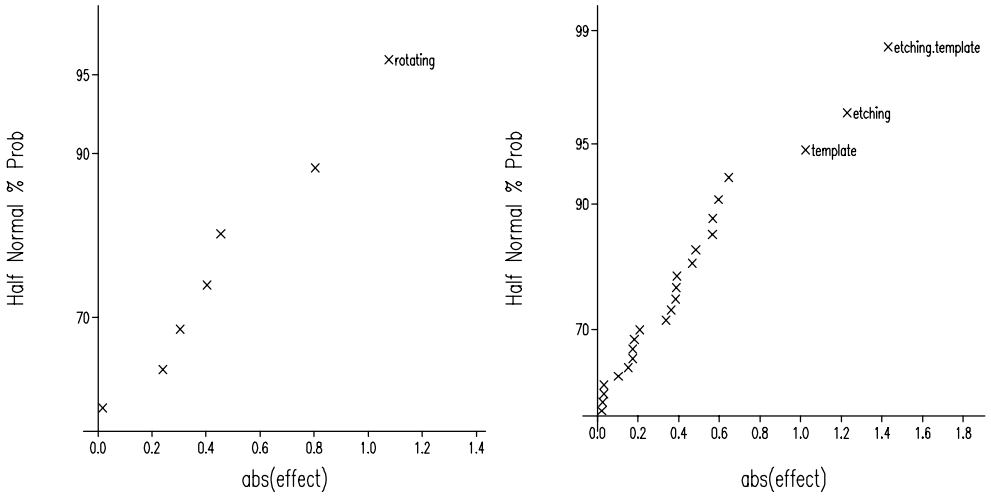


Fig. 3. Half-normal plots of Yates effects for (a) the days stratum and (b) the gauze stratum of the chemical experiment

For two-level experiments with two nested error terms, B blocks (or W whole-plots) and a total of N observations Guideline 1 implies $\{B,W\} \geq 8$, and
$N / \{B,W\} \geq 2$ (the braces {} denote that either B or W applies to a particular experiment). Note that it could still make sense to use less than this number of blocks if we are prepared to sacrifice the between-blocks information.

The ability of the halfnormal plots to separate active contrasts from inactive contrasts may be improved by manipulating the division of the contrasts over the distinct error strata. The required manipulation can be achieved through changes in experimental conduct. An example will now be given; refer to Table 9 throughout. In the chemical experiment, a factor called 'etching' specifies whether gauzes are treated in a jar with diluted acid, or kept stored in a jar with alcohol. If we wish, on the one hand, to evaluate the effect of etching within the blocks, while, on the other hand, we have one jar available for each alternative factor setting, then we need the following model.

$$u_{ij} = \beta_0 + x_1^T \beta_1 + e_i + x_3^T \beta_3 + e_{p(i)} + x_2^T \beta_2 + e_{j(ip)} \qquad (2)$$

Here, u_{ijp} is the respons of gauze j in jar p of block i, $\boldsymbol{x_1}$ a column vector with the B -1 efects between the blocks, $\boldsymbol{x_3}$ a column vector of B effects consisting of 'etching' and all generalized interactions of this factor with the between-blocks effects and $\boldsymbol{x_2}$ a column vector with N - $2B$ remaining effects. The corresponding true effects of the explanatory variables are denoted with β_1, β_3, and β_2, respectively, while β_0 denotes the general mean. The terms e_i, $e_{p(i)}$, and $e_{j(ip)}$ are contributions due to random differences between the blocks, between the jars within a block and between the gauzes put in one and the same jar, respectively. These correspond to three error strata, designated the day, jar, and gauze strata, respectively.

The day stratum, here, is a nesting stratum for the jars stratum. The jar stratum can be 'dissolved' into the day stratum by using a single jar per day and making etching a between-days factor. This results in an alternative division of the effects over two distinct strata. Such a division may also be accomplished by dissolving the jar stratum into the nested gauze stratum, through the simultaneous use of four jars per day. In the first case, the error between days is inflated, while the error to be inflated in the second case is that between gauzes. In either case, the index p in formula (2) is not needed any more. We

Table 9. Alternative divisions of the number of contrasts over the distinct error strata of the chemical experiment

no. jars / day	distinct error strata	no. contrasts / stratum
2	day, jar, gauze	B-1, B, N-2B
1	day+jar, gauze	B-1, N-B
4	day, jar+gauze	B-1, N-B

formalise the concept somewhat by giving the following.

Guideline 2. A nested stratum can be dissolved into a nesting stratum by conducting the experiment such that the index needed for the nested stratum equals 1 for all runs. A nesting stratum can be dissolved into a nested stratum by conducting the experiment such that the index for the nesting stratum equals that for the nested stratum for all runs.

The alternative divisions of the contrasts decrease the necessary number of halfnormal plots, and increases the number of contrasts in one of the remaining plots. This may improve the ability of the plot to separate active from inactive effects.

Finding treatment generators under a predetermined error structure

If a design for the treatment factors is incompatible with an intended split-plot error structure, then either the number of whole plots and subplots have to be adapted to the given set of treatments or the set of treatments have to be adapted to the given error structure. The case of adapting the number of plots can be solved with standard techniques. For the second case we propose the following.

Guideline 3. A predetermined split-plot error structure may be matched with a set of treatments by making separate subdesigns for the factors to be varied in each of the error strata, and merging of the subdesigns into an overall design.

Guideline 3 will now be used to derive a heuristic step-by-step procedure. The procedure is summarised in Table 10. Each step will be applied immediately to the cheese making

Table 10. Construction of a fractional split-plot design with given number of whole plots, subplots, whole plot factors, and subplot factors

lower stratum	upper stratum
(1) fraction	(5) fraction
(2) number of blocking generators	
(3) factor-settings and blocks	(6) factor-settings and blocks
(4) lower effect set between blocks	(7) upper effect set between blocks
linking of subdesigns	
(8) reduced lower and upper sets (aliases removed)	
(9) linking effects in lower and upper sets	

experiment.

(1) Determine the fraction of the subdesign for the *s* factors varied in the lowest error stratum. With a total of $N = 2^n$ units this will be a $2^{s-(s-n)}$ design.

For the cheese making experiment, we have a total of 2^7 units and 4 factors to be varied within the curds productions. We thus have a $2^{4-(4-7)}$ subdesign, or a 2^4 in 8 replicates.

(2) Determine the number *g* of blocking generators for the subdesign that involves the factors in the lowest error stratum. For this subdesign we have $B = 2^b$ blocks of 2^k units to make up a $2^{s-[(s-k)-b]}$ design. Each block makes up a $2^{s-(s-k)}$ design. If $(s-k) \leq 0$, then each block has all the treatments at least once. Therefore $g = 0$. If $(s-k) > 0$, it should be determined whether the subdesign is replicated at least once, i.e., whether or not $(s-k) \leq b$. If $(s-k) \leq b$, $g = s-k$, if $(s-k) > b$, $g = b$. It follows that $g = \max\{0, \min\{b, s-k\}\}$. This is also the number of generators for the overall experiment additional to those needed for the respective subdesigns.

The cheese-making experiment has $b=5$ and $s-k=2$; we thus need two blocking generators. These are to be equated to effects defined by the factors varied between the curds to link the subdesigns.

(3) Find settings for the factors varied in the lowest error stratum, given the fraction, resulting in a defining relation of minimum aberration. Find *g* blocking generators, given the MA design, resulting in an MA word-length pattern of the effects confounded with blocks. If, however, there are main effects that cannot be estimated within the blocks, use the fractional design that is second best, or even third best, according to the aberration criterion. Sun *et al*. (1997) published an extensive catalogue of blocking schemes of full and fractional factorial designs, which can be readily used for the purpose.

(4) Determine the 'lower effect set between blocks' as the set of blocking generators and all products of two up to *g* generators.

For the cheese making experiment, we have to block a full 2^4 design in four incomplete blocks of four cheeses. Let H, J, K, and L, respectively, denote the 'cheese' factors (note the skipping of 'I' as a factor label). A standard way to block is using HJK and JKL as blocking generators. Their generalized interaction, HL, is also confounded. Thus, we arrive at the lower effect set between blocks {HJK, JKL,HL}.

(5) Determine the fraction of the subdesign for the *w* factors varied in the upper stratum with $B = 2^b$ units. This will be a $2^{w-(w-b)}$ design.

For ease of exposition, we will jointly consider the 'milk' and 'curds' stratum of the cheese-making experiment as the upper stratum. We have a total of 32 curds productions and seven factors to be varied between the curds productions. Therefore, we have to look for a 2^{7-2} design.

(6) Find an MA design for the factors in the upper error stratum, and g blocking generators, analogous to step (3) for the lower stratum factors.

(7) Find the 'upper effect set between blocks' of 2^g -1 effects in the subdesign of the upper stratum, defined by the blocking generators of step (6). These are to be linked with those in the blocking set.

For the cheese making experiment the MA design has defining relation {I, ABCDF, ABDEG, CEFG} (Box *et al.*, 1978). We used ACG+BDFG+BCDE+AEF and BCG+ADFG+ACDE+BEF as generators for the effect set. Their generalized interaction is AB+CDF+DEG+ABCEFG. Thus, the upper effect set has a word-length pattern of one two-letter word, six three-letter words, four four-letter words, and one six-letter word. The catalogue of Sun *et al*. (1997) uses other generators, but the word length pattern in the resulting set is identical to our one. We conclude that we are effectively using an MA upper effect set.

(8) Make reduced upper and lower effect sets by retaining an effect of lowest order of each of a series of aliased effects.

The lower effect set of the cheese making experiment needs no reduction because it has no aliased effects. The upper effect set is reduced to {AB,ACG,BCG}.

(9) We will now obtain the g generators for the overall design additional to those of the respective subdesigns. For g = 1, join the effects in the reduced upper and lower effect sets. For g = 2, join an effect of highest order in the reduced upper effect set with one of lowest order in the reduced lower effect set. The second additional generator is formed by joining an effect of highest order from the reduced lower effect set with one of lowest order from the reduced upper effect set. For g > 2, apply the procedure for g=2, remove the effects in the reduced effect sets already spent, plus their generalised interaction, and repeat the procedure until exhaustion. The final set of linked effects will be called 'linking set'.

In our cheese-making experiment, the AB interaction of the upper effect set is equated to a second-order interaction between the cheese factors, JKL, say. Reversely, the HL interaction of the lower effect set is equated to a second-order interaction between the curds factors, ACG, say. Thus, we arrive at the linking set {ACGHL, BCGHJK, ABJKL}.

The generators found in (9) effectuate the merging of both subdesigns. At this stage, all blocking generators and treatment generators required to define the experiment are found. In a subsequent stage, the word-length pattern of the overall design should be derived. A quick appraisal can be obtained by looking at the separate patterns for the respective subdesigns and for the linking set. Indeed, it can be shown that the product of any word in the linking set of length x with the words in the defining relations of either subdesign has a length at least equal to x. Therefore, the resolution of the overall design can easily be found.

For the cheese making experiment, the words in the linking set are of length five, five, and six, respectively. The subdesign of the upper stratum has a defining relation with word-

lengths 4, 5, and 5, respectively. The lower stratum is not a fraction, so has no word-length pattern. We conclude that the joint design has a resolution of IV. This is the maximum resolution attainable in view of the subdesign for the curds.

Completing the designs

Chemical experiment

In the chemical experiment, six factors could potentially introduce error strata additional to the day and the gauze stratum (refer to Table 8 for a list of factors and their levels). Three of the relevant factors, viz., 'shaking time' (of the mixture), 'aging' (of the mixture after shaking), and 'time (before opening the autoclaves) after synthesis' lead to groups of experimental units treated during the same amount of time. Taken within blocks, these factors would lead to additional error strata. In keeping with our strategy, they should be varied between days.

The heating of the autoclaves was performed in an oven either standing at rest or under rotation according to the levels of the factor 'rotating'. The same oven could not perform both tasks simultaneously. Because it was impractical to use more than two ovens, this factor was clearly a between-days factor.

The factors 'etching' and 'pretreatment' require treatments to be applied in jars, either singly or in groups. Keeping their main effects within the blocks is preferred, because we already have to accommodate four other main effects between the blocks. To avoid an additional jar stratum, we use separate jars for each gauze (see Table 9).

The 4 x 8 blocking option cannot have the main effects for the three temporal factors and the factor rotating between the blocks. Putting these factors within the blocks would introduce four error strata of four degrees of freedom (df) each, additional to the between-blocks stratum (3 df) and the lowest error-stratum (12 df). This is a very problematic arrangement, which is not to be used.

Preliminary treatment combinations for the 8 x 4 arrangement were created through folding over of a 2^{15-11} design. The generators of the resulting design were ABCF, ABDG, ACDH, BCDJ, ABEK, ACEL, ADEM, BCEN, BDEO, CDEP, ACOQ (note the absence of 'I' as factor-label). Each word of this list can give the four factors to be put between the 8 blocks. Thus, an 8 x 4 arrangement is satisfactory where a 4 x 8 arrangement is not.

The design of the treatments was concluded by, first, attributing the coding letters A, B, C, and F to the factors to be put between the blocks. Secondly, four-level factors were defined using the sets {D,E} and {G,P} for the factors Si-source and Al-source, respectively. Finally, the remaining letters were arbitrarily allocated to the remaining factors. As the 'interactions' DE and GP are actually part of the main effects of the four-level factors, this is a resolution III design.

The weight gains of the 32 gauzes are given in Table 11. Halfnormal plots of log-transformed data, given in Fig. 3, revealed main effects of the factors rotating, etching, and template; there also appears to be a large interaction of etching an template. This interaction is actually aliased with seven other interactions. One of these is the interaction between rotating and cooling; the others involve factors without apparent main effects. Most likely, therefore, the significant effect results from the etching x template interaction.

Table 11. Data from the chemical experiment[a]

0.548	1.829	0.342	1.165	0.110	1.420	1.834	0.013
2.094	0.165	1.958	1.220	0.859	0.987	0.058	1.669
1.673	10.692	0.273	1.821	0.045	-0.05[b]	1.021	1.737
5.390	5.057	1.647	0.407	0.100	1.671	1.774	4.703

[a] Weight gain in Yates Order for factors A, B, C, D, E (factor A varying fastest), from left to right. See Table 8 for factors and their levels. [b] Replaced by +0.05 in subsequent analyses.

Mean squares of the active effects, and of the pooled remaining effects in each of the strata, are given in Table 12. The remainders in both strata are comparable in size. This suggests that the safeguard provided by the blocking was not necessary *a posteriori*. However, one should not rashly abandon applying the safeguard, especially where it is compatible with experimental practice.

Cheese-Making Experiment

The milk stratum of the cheese-making experiment was defined with generators A, B, and ACE, respectively. The four-level factor in the cheese stratum should not be constructed with factors H and L, because of the confounding of the HL interaction. We chose factors H and J instead. Counting occurrences of 'HJ' as one letter (Wu and Zhang, 1993), the defining relation of the final design has one word of four letters, seven words of five letters, six words of six letters, and one word of nine letters. As the word with four letters is

Table 12. Mean squares for effects on weight gain in the chemical experiment[a]

stratum	source of variation	df	mean square
between days	rotating	1	9.273
	other effects	6	1.556
within days	etching	1	12.076
	template	1	8.384
	etching x template	1	16.365
	other effects	21	1.046

[a] log-transformed data from Table 11.

inevitable in view of the maximum number of curds productions, we conclude that the number of distinguishable two-factor interactions in the overall design is maximized. More details of the design, and an analysis of a coded response variable are given in Schoen (1997).

Discussion

The present study motivates the confounding of main effects with blocks from the data-analytic consequences of using whole-plot factors within blocks. In the chemical experiment, we avoided extra error strata by putting the main effects of 4 factors between the blocks. This decided the issue of choosing between a 4 x 8 and 8 x 4 arrangement in favour of the second arrangement.

The cheese-making experiment is of the split-plot type. The treatment combinations had to be fitted into a predetermined block structure enforced by the process and by the experimental budget. Addelman (1964) gives a partial catalogue to construct the treatments for split-plot confounded designs, but does not give a solution to the problem at hand. Therefore, we had to construct the design ourselves. Because Addelman (1964) does not give details on his construction principles, it is necessary there to make these more explicit. This partly motivated this study.

The construction principles proposed here involve the construction of lower and upper effect sets between blocks. The problem of finding optimal generators of the upper effect set between blocks is analogous to the problem of finding those of the lower effect set, because in both cases we base the notion of optimality on the order of the effects in the set. In the case of the lower effect set, the effects are to be confounded with the blocks, while in the case of the upper effect set, the generators determine the confounding of upper stratum effects with those of the lower stratum.

The proposed subdesigns are from the catalogue of blocked designs of Sun *et al*. (1997). They have a MA word-length pattern of the lower effect set and the upper effect set. This heuristic approach is believed to lead to acceptable designs in terms of word-length pattern of the overall defining relation. However, it is possible that designs with lesser overall aberration may be constructed from subdesigns of higher aberration than their MA alternatives. Clearly, a systematic approach is needed to develop optimal blocking schemes when main effects are to be confounded with blocks.

The approach of Bingham and Sitter (1999) systematically generates such designs. However, in contrast with our procedure, this approach cannot lead to designs with replicated whole plots when the whole plot factors alone are considered. These designs may be needed for the following operational reasons. First, if the whole plot effects have to be assessed statistically, a minimum of eight whole plots will be necessary. Second, designs with replicated settings of the whole plot factors will have twice the number of effects of a corresponding design with unreplicated settings of whole plot factors. Thus, the replicated designs will be more informative. We conclude that more work is needed to create MA two-level split-plot designs with replicated settings of the whole plot factors.

An alternative for the construction and linkage of separate subdesigns is to cross the subdesign for the upper error stratum in, say, w factors with a full factorial two-level design with, say, q within-blocks factors. Each of the remaining within-blocks factors can be

added to the design with a generator containing the new factor, at least one of the q factors, and at least one of the *w* factors. This procedure guarantees that the main effects of the added factors are within the blocks. Generators with even word-lengths give designs having a resolution of at least IV. The aforementioned work of Bingham and Sitter (1999) connects this procedure to notions of optimality.

Miller (1997) discussed fractional designs with crossed error structures. He proposes to construct separate 'between rows' and 'between columns' subdesigns. Fundamental to his technique is the finding of two sets of blocking generators, one for each subdesign, and equating the generators to each other. Thus, his construction method resembles that proposed here for nested error structures, because we also have to find two sets of generators. The difference between the two approaches, induced by the difference between crossed and nested designs, is that one of our sets of generators is not really involved in blocking, but rather in linking of the subdesigns.

Bisgaard (1994a, 1994b) discussed the implications on the design resolution of potential interactions between blocks, treated as a fixed factor, and experimental factors. Indeed, it may well be argued that blocks that are physically different should be handled as a fixed factor rather than as a random one. In our examples, however, the sets of treatments were randomly attributed to each of the blocks, while there were no apparent physical differences other than those induced by the treatments. Therefore, we think that the assumption of a random between-blocks effect having no interactions with the experimental factors is reasonable here.

Chapter 3 Analysis of unreplicated designs

3.1 Three robust scale estimators to judge unreplicated experiments

Introduction

Results from unreplicated experiments are commonly evaluated by making halfnormal or normal plots of standardised single degree of freedom contrasts. Daniel (1959) introduced halfnormal plotting. He plotted absolute values of standardised contrasts from two-level experiments against expected quantiles of the halfnormal distribution. If the experimental error is normal, and if there are only a few active contrasts, then most will come from a normal distribution with zero mean and the same unknown variance. Inactive contrasts will then roughly lie on a straight line through the origin in the plot. Those not compatible with this line are designated active. Thus, halfnormal plots will separate the few active contrasts from the inactive ones.

Daniel (1976) pointed out that plotting signed contrasts against expected quantiles of the normal distribution will also separate active from inactive contrasts. Because such plots are more capable of detecting anomalies in the data, Daniel (1976) recommends full normal plots.

The subjectivity as to which contrasts deviate from the straight line is a well-known problem with (half)normal plotting. To overcome this problem, various robust estimators of the contrasts' standard errors have been developed; see Haaland and O'Connell (1995) for an overview. Haaland and O'Connell (HOC) define a family of estimators constructed from the ordered absolute values of the contrasts, by the following common elements. First, an initial estimate of the standard error is calculated as a quantile of the full set of absolute valued contrasts, multiplied by a consistency constant determined from the normal distribution. Second, potential active contrasts are stripped from the others by retaining only those smaller than a constant times the initial estimate. Third, a scalar function of the remaining contrasts is multiplied with a simulated consistency constant to give the final estimate of the standard error. Based on simulation results for 15 contrasts, HOC recommend three estimators according to *a priori* ideas on the likely number of active contrasts (1-3, 4-6, and 7-8, respectively). One of the three recommended estimators is based on the median of the full set and the root mean square of the retained set of contrasts. This so-called Adaptive Standard Error (ASE) was originally proposed by Dong (1993). The other two estimators are based on the median or the 0.45 quantile of the full set, respectively, and the median of the retained contrasts; these are Pseudo Standard Errors (Lenth, 1989); the two versions will be designated PSE(50) and PSE(45), respectively. In general, the ASE is less robust against contamination with active contrasts than the PSE(50), because it uses all the contrasts below the cut-off point. The PSE(50) is obviously less robust than the PSE(45).

HOC suggest judging ratios of the contrasts and the estimated standard error against critical values determined by simulation. Their paper has consistency constants to calculate

ASE and PSE(50), and critical values to test with the PSE(50), each for experiments with 7, 11, 15, 17, 23 and 31 contrasts. The present study extends the results to all contrast numbers in the range from 7 to 127, and all three estimators. It also offers a guideline for choosing among the estimators. The tables in this study show the results for contrast numbers from 7 to 96, 112, 120, 124, and 127. The remaining results can be obtained from the author.

The extension of the HOC results to all contrast numbers in the chosen range is motivated by the great variety of possible designs. Two-level designs with randomisation restrictions, for example, have their contrasts divided over sets of different precision. Specifically, split-plot and strip-block arrangements (Box and Jones, 1992) result in a division over two and three sets of contrasts, respectively, while an extension of the latter arrangement (Miller, 1997) has a division into four sets of contrasts. The contrast numbers of at least one of the sets will not equal 2^n-1. The examples from Study 2.1 and Study 2.2 are cases in point. Further, the Plackett-Burman designs have contrast numbers that equal $4p$ - 1, p being an integer ranging from 2 up to 25. HOC do not cover contrast numbers 19 and 27, or any number above 31. Finally, when factors other than two-level ones are included, the total number of contrasts may equal neither 2^n - 1 nor $4p$ - 1. More specifically, Box and Draper (1987), and Draper *et al*. (1994) illustrate the break down of results from central composite and Box-Behnken experiments, respectively, into single degree of freedom contrasts.

We extended the HOC results to a wider range because it is not uncommon to have experiments with larger numbers of contrasts than 31. The Box-Behnken design of Draper *et al*. (1994), for example, has 46 runs; Schoen (1997) used a 128-run two-level split-split-plot experiment to investigate the effect of 11 factors on properties of Dutch cheeses.

The remaining part of this study is organised as follows. First, we define the estimators used, and present the method of calculating the required constants. Subsequently, we present consistency constants and critical values for testing of the contrasts with Type I errors of 0.20, 0.15, 0.10, 0.05, and 0.01, respectively. Further, we propose a guideline for choosing among the three estimators. A practical example concludes the study.

Definitions and methods

The estimators of the standard error used in this study are designated ASE, PSE(50) and PSE(45), respectively. They are based on the set of absolute values of the contrasts, designated {| c_j | ; j = 1...k}, with c_j denoting the contrast with ordered absolute value number j, and k denoting the total number of contrasts in the set. Their calculation involves (1) forming of an initial estimate, (2) trimming of contrasts being large relative to the intial estimate, and (3) forming of the final estimate.

The initial estimate (Step 1) is calculated as

$$s_{ini} = cc_1 * quantile(q; \{|c_j|,\ j = 1,2,...,k\})$$

Here, q is 0.5 for ASE and PSE(50), or 0.45 for PSE(45), and cc_1 is a consistency constant, defined as

$$cc_1 = \frac{1}{\Phi^{-1}\left(\left[1 + (i - 0.5)/k\right]/2\right)}$$

with Φ(.) denoting the cumulative standard normal distribution function. For ASE and PSE(50), *i* is either an average number (even *k*), or a single number (odd *k*), of the contrast(s) involved in establishing the median. Thus the quantile effectively used for ASE and PSE(50), that is, (*i* - 0.5)/*k*, equals 0.5 throughout.

For PSE(45), *i* is either an average number (for *k* equaling a multiple of 20), or a single number (for other values of *k*), of the contrast(s) involved in establishing the 0.45 quantile. The values of *i* are given in Table 15, to be discussed later. Note that the quantile effectively used for PSE(45) varies with *k*. The alternative, using (*i* - 0.5)/*k* = 0.45 throughout, is not considered, because this would result in a biased initial estimate even if there were no active contrasts.

In the trimming stage (Step 2), contrasts larger than $b \,.\, s_{ini}$ are stripped from the full set of contrasts. The constants *b* used here are 2.5 for ASE and PSE(50), and 1.25 for PSE(45). These values are based on HOC results on the power of detecting active contrasts for the case *k*=15.

To form the final estimate (Step 3), the root mean square of the retained contrasts, for the ASE, or its median, for both versions of the PSE, is multiplied with a consistency constant designated cc_2. The constants are obtained as follows. For each number of contrasts, 10 000 datasets were generated from a standard normal distribution. The sets were stripped from the largest contrasts according to the specifications of the estimators. From the remaining contrasts in each set, either their root mean square, (ASE), or their median (PSE) was determined. The reciprocal of the mean of these quantities over all sets is the required cc_2.

Critical values of tests with the above estimators were obtained using 10 000 sets of *k* contrasts from a standard normal distribution for each of the estimators, with *k* ranging from 7 to 127. For each set, ASE, PSE(45), or PSE(50), was calculated using the simulated value for cc_2. For each contrast in the set, the absolute value was divided by this estimate. Critical values for double-sided tests, for type I errors of 0.20,0.15, 0.10, 0.05, and 0.01, respectively, were determined as quantiles of the empirical distribution of all *k* * 10 000 absolute-valued contrasts standardised with their estimated standard error. Thus, we use an effect-wise error rate.

Results

Tables

For each number of contrasts (*k*) considered, simulated consistency constants (cc_2) for calculation of the final scale estimators, and the critical values for the testing of contrasts are given in Tables 1, 2, and 3, for ASE, PSE(50), and PSE (45), respectively.

Table 13. Consistency constants and critical values for ASE

initial cutoff: 3.707 x {median of absolute contrasts}						
		Type I error				
k	cc_2	0.20	0.15	0.10	0.05	0.01
7	1.133	1.26	1.39	1.54	1.79	4.53
8	1.115	1.26	1.39	1.55	1.79	3.90
9	1.120	1.26	1.40	1.57	1.84	3.81
10	1.106	1.26	1.40	1.57	1.83	3.44
11	1.104	1.27	1.42	1.60	1.88	3.49
12	1.104	1.27	1.41	1.59	1.86	3.29
13	1.101	1.27	1.42	1.60	1.88	3.29
14	1.095	1.27	1.42	1.60	1.88	3.13
15	1.095	1.27	1.41	1.60	1.89	3.10
16	1.089	1.27	1.42	1.61	1.89	3.07
17	1.090	1.27	1.42	1.61	1.90	3.05
18	1.083	1.28	1.43	1.62	1.91	3.01
19	1.088	1.27	1.42	1.62	1.90	2.94
20	1.078	1.28	1.43	1.62	1.91	2.93
21	1.078	1.28	1.43	1.62	1.91	2.92
22	1.076	1.28	1.43	1.63	1.92	2.91
23	1.080	1.28	1.43	1.62	1.92	2.89
24	1.076	1.28	1.43	1.62	1.92	2.85
25	1.077	1.28	1.43	1.63	1.93	2.85
26	1.075	1.28	1.43	1.62	1.92	2.81
27	1.076	1.27	1.43	1.62	1.92	2.81
28	1.072	1.28	1.43	1.63	1.93	2.81
29	1.072	1.28	1.44	1.63	1.93	2.82
30	1.069	1.28	1.44	1.63	1.93	2.79
31	1.072	1.28	1.43	1.63	1.93	2.78
32	1.071	1.28	1.43	1.63	1.93	2.76
33	1.069	1.28	1.43	1.63	1.93	2.76
34	1.071	1.28	1.43	1.63	1.92	2.75
35	1.069	1.28	1.43	1.63	1.94	2.74
36	1.066	1.28	1.44	1.64	1.94	2.74
37	1.068	1.28	1.43	1.63	1.93	2.73
38	1.065	1.28	1.43	1.64	1.94	2.72
39	1.066	1.28	1.43	1.63	1.94	2.72
40	1.067	1.28	1.43	1.64	1.94	2.73
41	1.065	1.28	1.43	1.64	1.94	2.72
42	1.067	1.27	1.43	1.63	1.93	2.70
43	1.064	1.28	1.44	1.63	1.94	2.71
44	1.064	1.28	1.44	1.64	1.95	2.72
45	1.063	1.28	1.44	1.64	1.94	2.70
46	1.065	1.28	1.43	1.63	1.94	2.69
47	1.063	1.28	1.44	1.64	1.95	2.71
48	1.065	1.28	1.43	1.63	1.94	2.68
49	1.063	1.28	1.43	1.64	1.94	2.68
50	1.064	1.28	1.43	1.63	1.94	2.67
51	1.063	1.28	1.43	1.64	1.94	2.66
52	1.063	1.28	1.43	1.63	1.94	2.68
53	1.063	1.28	1.43	1.63	1.94	2.66

Table 13. (continued)

		Type I error				
k	cc_2	0.20	0.15	0.10	0.05	0.01
54	1.061	1.28	1.43	1.64	1.94	2.67
55	1.063	1.28	1.43	1.63	1.93	2.65
56	1.061	1.28	1.43	1.64	1.94	2.67
57	1.061	1.28	1.44	1.64	1.94	2.66
58	1.061	1.28	1.43	1.64	1.94	2.67
59	1.060	1.28	1.44	1.64	1.94	2.66
60	1.062	1.28	1.43	1.63	1.94	2.65
61	1.058	1.28	1.44	1.64	1.95	2.66
62	1.061	1.28	1.43	1.64	1.94	2.65
63	1.061	1.28	1.44	1.64	1.94	2.64
64	1.059	1.28	1.44	1.64	1.95	2.66
65	1.062	1.28	1.43	1.63	1.94	2.63
66	1.060	1.28	1.44	1.64	1.94	2.65
67	1.059	1.28	1.44	1.64	1.95	2.65
68	1.058	1.28	1.44	1.64	1.95	2.65
69	1.058	1.28	1.44	1.64	1.95	2.65
70	1.057	1.28	1.44	1.64	1.95	2.65
71	1.058	1.28	1.44	1.64	1.95	2.64
72	1.057	1.28	1.44	1.64	1.95	2.65
73	1.060	1.28	1.43	1.64	1.94	2.64
74	1.058	1.28	1.44	1.64	1.95	2.64
75	1.056	1.28	1.44	1.64	1.95	2.64
76	1.056	1.28	1.44	1.64	1.95	2.65
77	1.057	1.28	1.44	1.64	1.95	2.64
78	1.057	1.28	1.44	1.64	1.95	2.63
79	1.058	1.28	1.44	1.64	1.95	2.63
80	1.058	1.28	1.44	1.64	1.94	2.64
81	1.056	1.28	1.44	1.64	1.95	2.65
82	1.055	1.28	1.44	1.64	1.95	2.63
83	1.054	1.28	1.44	1.64	1.95	2.64
84	1.056	1.28	1.44	1.64	1.95	2.63
85	1.056	1.28	1.44	1.64	1.95	2.63
86	1.058	1.28	1.43	1.64	1.95	2.63
87	1.055	1.28	1.44	1.64	1.96	2.64
88	1.057	1.28	1.44	1.64	1.94	2.62
89	1.056	1.28	1.44	1.64	1.95	2.62
90	1.056	1.28	1.44	1.64	1.95	2.63
91	1.056	1.28	1.44	1.64	1.95	2.62
92	1.056	1.28	1.44	1.64	1.95	2.62
93	1.056	1.28	1.44	1.64	1.95	2.63
94	1.056	1.28	1.44	1.64	1.95	2.63
95	1.056	1.28	1.44	1.64	1.95	2.62
96	1.056	1.28	1.44	1.64	1.95	2.64
112	1.054	1.28	1.44	1.64	1.95	2.61
120	1.055	1.28	1.44	1.64	1.95	2.61
124	1.055	1.28	1.44	1.64	1.95	2.61
127	1.054	1.28	1.44	1.64	1.95	2.61

Table 14. Consistency constants and critical values for PSE(50)

initial cutoff: 3.707 x {median of absolute contrasts}

		Type I error				
k	cc_2	0.20	0.15	0.10	0.05	0.01
7	1.446	1.25	1.47	1.79	2.40	5.25
8	1.457	1.24	1.43	1.71	2.25	4.90
9	1.465	1.26	1.46	1.75	2.30	4.59
10	1.469	1.25	1.44	1.70	2.21	4.26
11	1.463	1.28	1.48	1.77	2.30	4.20
12	1.486	1.26	1.45	1.72	2.20	4.01
13	1.479	1.27	1.47	1.74	2.23	3.99
14	1.484	1.26	1.45	1.71	2.18	3.79
15	1.485	1.27	1.45	1.71	2.16	3.65
16	1.484	1.27	1.46	1.71	2.17	3.61
17	1.487	1.27	1.46	1.72	2.16	3.54
18	1.488	1.28	1.46	1.72	2.15	3.50
19	1.495	1.27	1.45	1.70	2.13	3.40
20	1.485	1.28	1.46	1.71	2.14	3.39
21	1.481	1.28	1.46	1.71	2.13	3.34
22	1.482	1.29	1.47	1.72	2.13	3.35
23	1.491	1.28	1.46	1.71	2.12	3.27
24	1.489	1.28	1.46	1.71	2.11	3.25
25	1.491	1.28	1.46	1.71	2.12	3.23
26	1.493	1.28	1.45	1.70	2.09	3.15
27	1.494	1.28	1.45	1.69	2.08	3.14
28	1.496	1.28	1.45	1.69	2.08	3.14
29	1.493	1.29	1.46	1.70	2.09	3.14
30	1.489	1.29	1.47	1.70	2.10	3.10
31	1.493	1.28	1.45	1.69	2.07	3.08
32	1.494	1.28	1.46	1.69	2.08	3.06
33	1.493	1.28	1.45	1.68	2.07	3.03
34	1.499	1.28	1.45	1.68	2.06	3.02
35	1.489	1.29	1.46	1.69	2.07	3.02
36	1.496	1.28	1.46	1.69	2.06	3.00
37	1.497	1.28	1.45	1.67	2.05	2.97
38	1.497	1.28	1.45	1.68	2.05	2.95
39	1.492	1.29	1.46	1.69	2.06	2.97
40	1.501	1.28	1.46	1.69	2.06	2.97
41	1.496	1.28	1.45	1.68	2.06	2.95
42	1.501	1.27	1.44	1.67	2.04	2.91
43	1.494	1.28	1.45	1.68	2.05	2.93
44	1.498	1.29	1.46	1.69	2.06	2.93
45	1.495	1.28	1.45	1.68	2.04	2.91
46	1.500	1.28	1.45	1.68	2.04	2.88
47	1.492	1.29	1.46	1.69	2.06	2.93
48	1.501	1.28	1.45	1.67	2.03	2.87
49	1.495	1.28	1.45	1.68	2.04	2.89
50	1.502	1.28	1.45	1.67	2.03	2.86
51	1.497	1.28	1.45	1.67	2.03	2.85
52	1.502	1.28	1.44	1.67	2.02	2.85
53	1.501	1.28	1.45	1.67	2.03	2.83

Table 14. (continued)

		Type I error				
k	cc_2	0.20	0.15	0.10	0.05	0.01
54	1.496	1.28	1.45	1.67	2.03	2.85
55	1.500	1.28	1.44	1.66	2.01	2.82
56	1.500	1.28	1.45	1.67	2.02	2.84
57	1.500	1.28	1.45	1.67	2.03	2.83
58	1.504	1.28	1.44	1.66	2.02	2.82
59	1.500	1.28	1.44	1.67	2.01	2.81
60	1.503	1.28	1.45	1.67	2.02	2.80
61	1.496	1.29	1.45	1.67	2.03	2.83
62	1.502	1.28	1.45	1.67	2.02	2.81
63	1.501	1.28	1.45	1.67	2.02	2.80
64	1.499	1.29	1.45	1.67	2.02	2.82
65	1.502	1.28	1.45	1.67	2.01	2.79
66	1.502	1.28	1.45	1.66	2.01	2.79
67	1.499	1.28	1.45	1.67	2.02	2.79
68	1.502	1.28	1.45	1.67	2.02	2.79
69	1.495	1.29	1.46	1.68	2.03	2.80
70	1.498	1.29	1.45	1.67	2.02	2.79
71	1.499	1.29	1.45	1.67	2.02	2.78
72	1.498	1.29	1.45	1.67	2.02	2.79
73	1.504	1.28	1.44	1.66	2.00	2.77
74	1.499	1.28	1.45	1.67	2.02	2.77
75	1.495	1.29	1.45	1.67	2.01	2.77
76	1.497	1.29	1.45	1.67	2.01	2.78
77	1.498	1.28	1.45	1.67	2.01	2.77
78	1.499	1.29	1.45	1.67	2.01	2.76
79	1.502	1.28	1.45	1.66	2.01	2.75
80	1.503	1.28	1.45	1.66	2.00	2.75
81	1.498	1.28	1.45	1.66	2.01	2.77
82	1.498	1.28	1.45	1.66	2.00	2.74
83	1.495	1.29	1.45	1.67	2.01	2.75
84	1.502	1.28	1.44	1.66	2.00	2.74
85	1.498	1.29	1.45	1.67	2.01	2.75
86	1.503	1.28	1.44	1.66	2.00	2.73
87	1.499	1.29	1.45	1.67	2.01	2.75
88	1.502	1.28	1.44	1.66	2.00	2.72
89	1.502	1.28	1.45	1.66	2.00	2.73
90	1.501	1.28	1.45	1.66	2.00	2.74
91	1.500	1.28	1.45	1.66	2.00	2.73
92	1.500	1.28	1.45	1.66	2.00	2.72
93	1.500	1.29	1.45	1.67	2.00	2.74
94	1.501	1.28	1.45	1.66	2.00	2.73
95	1.501	1.28	1.45	1.66	2.00	2.72
96	1.502	1.28	1.45	1.66	2.00	2.73
112	1.500	1.29	1.45	1.66	2.00	2.71
120	1.504	1.28	1.44	1.66	1.99	2.69
124	1.503	1.28	1.44	1.66	1.99	2.69
127	1.500	1.28	1.45	1.66	1.99	2.69

Table 15. Consistency constants and critical values for PSE(45)

	initial cutoff			Type I error				
k	contrast	multiplier	cc_2	0.20	0.15	0.10	0.05	0.01
7	4	1.853	1.837	1.46	1.81	2.33	3.39	6.84
8	4	2.158	1.725	1.34	1.67	2.16	3.13	6.63
9	5	1.853	1.843	1.42	1.72	2.15	3.04	6.02
10	5	2.091	1.745	1.36	1.67	2.13	3.01	5.87
11	5	2.325	1.812	1.36	1.65	2.08	2.90	5.72
12	6	2.048	1.762	1.36	1.64	2.06	2.86	5.33
13	6	2.241	1.806	1.36	1.64	2.05	2.81	5.32
14	7	2.018	1.778	1.35	1.61	1.99	2.69	4.71
15	7	2.182	1.800	1.36	1.62	2.01	2.70	4.78
16	8	1.996	1.797	1.35	1.61	1.98	2.64	4.58
17	8	2.138	1.808	1.36	1.61	1.99	2.66	4.55
18	9	1.980	1.806	1.34	1.59	1.94	2.57	4.26
19	9	2.105	1.814	1.34	1.59	1.93	2.54	4.20
20	9.5	2.091	1.835	1.34	1.58	1.92	2.52	4.10
21	10	2.079	1.815	1.35	1.59	1.93	2.53	4.13
22	10	2.190	1.846	1.35	1.58	1.92	2.50	4.05
23	11	2.057	1.829	1.34	1.57	1.90	2.46	3.89
24	11	2.158	1.859	1.34	1.56	1.88	2.43	3.92
25	12	2.040	1.845	1.33	1.56	1.86	2.40	3.72
26	12	2.132	1.837	1.35	1.58	1.89	2.43	3.83
27	13	2.025	1.841	1.33	1.55	1.85	2.36	3.68
28	13	2.110	1.841	1.34	1.56	1.87	2.40	3.72
29	14	2.012	1.835	1.34	1.56	1.85	2.36	3.58
30	14	2.091	1.851	1.34	1.55	1.85	2.37	3.69
31	14	2.170	1.864	1.33	1.55	1.85	2.35	3.60
32	15	2.075	1.863	1.33	1.54	1.84	2.33	3.54
33	15	2.148	1.867	1.33	1.55	1.84	2.33	3.49
34	16	2.061	1.861	1.34	1.55	1.84	2.32	3.49
35	16	2.129	1.862	1.32	1.53	1.82	2.29	3.43
36	17	2.048	1.859	1.33	1.54	1.82	2.30	3.43
37	17	2.113	1.868	1.32	1.53	1.81	2.27	3.42
38	18	2.037	1.871	1.33	1.53	1.81	2.26	3.37
39	18	2.098	1.872	1.33	1.54	1.81	2.27	3.35
40	18.5	2.091	1.863	1.33	1.54	1.81	2.27	3.34
41	19	2.085	1.880	1.31	1.52	1.79	2.24	3.33
42	19	2.142	1.880	1.32	1.52	1.78	2.23	3.26
43	20	2.073	1.877	1.33	1.53	1.80	2.24	3.29
44	20	2.127	1.882	1.32	1.52	1.78	2.22	3.25
45	21	2.062	1.877	1.32	1.52	1.79	2.23	3.24
46	21	2.114	1.877	1.32	1.52	1.79	2.23	3.25
47	22	2.053	1.877	1.33	1.53	1.80	2.24	3.27
48	22	2.102	1.895	1.31	1.51	1.77	2.19	3.19
49	23	2.044	1.884	1.31	1.50	1.76	2.18	3.15
50	23	2.091	1.886	1.32	1.51	1.77	2.21	3.21
51	23	2.138	1.883	1.32	1.52	1.78	2.20	3.19
52	24	2.081	1.888	1.31	1.50	1.76	2.18	3.14
53	24	2.126	1.882	1.33	1.52	1.78	2.21	3.19

Table 15. (continued)

	initial cutoff			Type I error				
k	contrast	multiplier	cc_2	0.20	0.15	0.10	0.05	0.01
54	25	2.072	1.890	1.32	1.51	1.77	2.19	3.16
55	25	2.115	1.888	1.31	1.51	1.76	2.18	3.15
56	26	2.063	1.884	1.32	1.51	1.77	2.18	3.13
57	26	2.105	1.894	1.31	1.50	1.75	2.15	3.08
58	27	2.055	1.895	1.31	1.50	1.75	2.16	3.07
59	27	2.096	1.894	1.31	1.50	1.76	2.17	3.09
60	27.5	2.091	1.894	1.30	1.49	1.74	2.14	3.04
61	28	2.087	1.894	1.31	1.50	1.75	2.15	3.07
62	28	2.125	1.897	1.31	1.51	1.76	2.16	3.07
63	29	2.079	1.898	1.31	1.50	1.75	2.15	3.06
64	29	2.116	1.902	1.31	1.49	1.74	2.14	3.04
65	30	2.071	1.888	1.31	1.50	1.75	2.15	3.05
66	30	2.107	1.896	1.31	1.49	1.74	2.14	3.03
67	31	2.064	1.892	1.30	1.49	1.73	2.12	3.00
68	31	2.099	1.895	1.31	1.49	1.74	2.13	3.00
69	32	2.057	1.899	1.31	1.50	1.74	2.14	3.02
70	32	2.091	1.899	1.31	1.50	1.74	2.13	3.01
71	32	2.125	1.894	1.31	1.50	1.74	2.13	3.00
72	33	2.084	1.897	1.31	1.49	1.74	2.13	2.98
73	33	2.117	1.900	1.31	1.50	1.74	2.13	3.02
74	34	2.077	1.899	1.30	1.49	1.73	2.12	2.97
75	34	2.109	1.897	1.31	1.50	1.74	2.13	2.99
76	35	2.071	1.905	1.31	1.49	1.73	2.12	2.97
77	35	2.101	1.896	1.31	1.50	1.74	2.13	3.00
78	36	2.065	1.901	1.30	1.49	1.72	2.11	2.94
79	36	2.094	1.902	1.31	1.50	1.74	2.12	2.97
80	36.5	2.091	1.892	1.31	1.50	1.74	2.12	2.96
81	37	2.088	1.904	1.30	1.49	1.73	2.11	2.94
82	37	2.117	1.905	1.30	1.49	1.72	2.11	2.95
83	38	2.082	1.902	1.30	1.49	1.72	2.10	2.93
84	38	2.110	1.905	1.31	1.49	1.72	2.10	2.93
85	39	2.076	1.903	1.31	1.49	1.73	2.11	2.93
86	39	2.103	1.905	1.30	1.49	1.72	2.10	2.93
87	40	2.070	1.892	1.31	1.50	1.74	2.12	2.95
88	40	2.097	1.906	1.30	1.49	1.72	2.10	2.92
89	41	2.065	1.907	1.30	1.48	1.72	2.09	2.91
90	41	2.091	1.906	1.31	1.49	1.72	2.10	2.91
91	41	2.117	1.905	1.30	1.49	1.72	2.09	2.91
92	42	2.085	1.904	1.30	1.48	1.71	2.09	2.90
93	42	2.111	1.908	1.30	1.48	1.72	2.09	2.91
94	43	2.080	1.914	1.30	1.48	1.71	2.08	2.88
95	43	2.105	1.906	1.31	1.49	1.72	2.09	2.89
96	44	2.075	1.909	1.30	1.48	1.71	2.09	2.87
112	51	2.086	1.910	1.30	1.48	1.70	2.07	2.84
120	54.5	2.091	1.912	1.30	1.47	1.70	2.06	2.83
124	56	2.104	1.909	1.30	1.48	1.71	2.07	2.82
127	58	2.077	1.913	1.30	1.47	1.70	2.06	2.81

The calculation of the median based estimators requires stripping of contrasts larger than 2.5 * s_{ini}. Thus the cutoff value is 2.5 / Φ^{-1} (0.75) = 3.707 times the median of the absolute contrasts.

The calculation of PSE(45) requires a cutoff of 1.25 * s_{ini}. This is not a fixed multiplier of the 0.45 quantile, in view of the quantile effectively used here (see the preceding section). To facilitate the use of PSE(45), we add in Table 15 both the number of the ordered absolute contrast and the multiplier needed to calculate the cutoff. Halves in the contrast number imply averaging the two contrasts with adjacent numbers.

Choosing among estimators

HOC recommend ASE, PSE(50) and PSE(45) for likely proportions of active contrasts of up to 0.2, between 0.2 and 0.4 and more than 0.4, respectively. These recommendations are based on power considerations for the case of 15 contrasts. For the contrast numbers studied in this study, we propose a roughly parallel recommendation based on the resolution of the design.

Consider, to begin with, a design of resolution V or higher. Such a design is often carried out when it is expected that all main effects and many of the interactions between two factors are active. For a design with f factors and 2^n runs, say, this implies up to $0.5\ f\ (f+1)$ active contrasts. An initial scale estimator based on the median breaks down when there are more than $0.5\ (2^n - 1)$ active contrasts, because it will be contaminated with some of these.

Table 16 gives particulars for 32-run, 64-run, and 128-run experiments with the maximum numbers of factors for a resolution V design. The potential number of inactive contrasts is calculated as the difference between the total number of contrasts and the potential number of active contrasts, $0.5\ f\ (f + 1)$. The medians correspond to contrast numbers much larger than the potential number of inactive contrasts, while the 0.45 quantile corresponds to a contrast number very near this potential number. Thus it is more reasonable to use the PSE(45) than the other two estimators.

Resolution IV designs are often used when we expect a good number of active main effects and possibly some interactions. An initial estimator based on the median seems reasonable here, because our expectation implies that about half of the contrasts could be active (there are resolution IV two-level designs for up to f factors in $2f$ runs). This number of contrasts being near the breakdown of the initial estimator, it seems wise to prefer PSE(50) to the ASE.

Finally, resolution III designs, in our view, are sensible only when a low proportion of the factors are expected to be active, because with an increasing number of active factors the number of potentially relevant interactions also increases. With a low proportion of active contrasts, we may as well choose a more efficient estimate than PSE(50); this makes a point

Table 16. Resolution V designs with maximum numbers of factors for 32, 64, and 128 runs

number of runs	number of factors	potential number of inactive contrasts	contrast number for 0.45 quantile
32	6	10	14
64	8	27	29
128	11	61	58

for the ASE in such cases.

The above argument relies on *a priori* views on the likely number of active contrasts. A more comprehensive approach would consider the power of the contrast test, using various numbers and sizes of the contrasts. Also, the tuning constant needed for the cutoff in the construction of the estimates may not be optimal when the total number of effects does not equal 15. A study covering these aspects is clearly beyond the scope of the present study. In the absence of such a study, we propose our argument as a rough but general way to choose among the estimators.

Example

Srinivas *et al*. (1994) use a Plackett-Burman design with 20 runs to select from 19 ingredients of the nutritional medium those enhancing the production of a certain enzyme by the fungus *Aspergillus niger*. The ingredients and their effects on enzyme activity after three days of fungus fermentation (expressed as units per gram dried biomass) are given in Table 17. The low and high levels of the factors correspond with less or more of the ingredients, respectively; see the original publication for more details. The problem at hand is to give a statistical assessment of the effects by testing against an estimate of their standard error.

We illustrate the calculations for all three estimators studied here, although we recommend the ASE in view of the resolution of the present design. For the median based estimators, the ordered absolute effect with ranking number 10, 6.29, is multiplied by 3.707 (obtained from Table 13) to get a cut-off value of 23.317. All effects are below this value,

Table 17. Ingredients and ordered contrast sizes for enzyme activity example

rank	ingredient	size
1	lactose	-1.25
2	sodium chloride	1.33
3	potassium chloride	1.57
4	magnesium sulphate	2.55
5	sodium nitrate	2.61
6	ammonium sulphate	-2.69
7	bengal gram flour	3.93
8	calcium chloride	-4.63
9	french bean flour	5.57
10	soy flour	6.29
11	diammonium hydrogen phosphate	-6.31
12	corn steep liquor	7.37
13	black gram flour	8.43
14	ferrous sulphate	-9.13
15	urea	10.47
16	ammonium nitrate	-10.71
17	guar flour	11.43
18	citric acid	11.55
19	ammonium chloride	-14.01

so all are retained for further calculation of the standard error estimates. The root mean square of these effects is 7.494; multiplying by 1.088 (obtained from Table 13) gives 8.153 as the ASE. The PSE(50) is obtained by multiplying the median of the retained effects, 6.29, with 1.495 (obtained from Table 14); this gives a value of 9.404.

The calculation of PSE(45) starts with establishing the multiplier and the ranking number of the ordered absolute effects needed to calculate the cutoff. These are given in Table 15 as 2.105 and 9, respectively. Multiplication of the 9th effect, 5.57, with 2.105 results in a cutoff of 11.725. There is one effect larger than this value. The final estimate uses the median of the 18 remaining effects, that is, (5.57+6.29)/2, multiplied by a consistency constant of 1.814. This results in a PSE(45) of 10.757.

In the original paper, a standard error of the effects is calculated on the basis of triplicated observations. Its value is about 3.37. This is considerably smaller than the robust estimates. A full discussion of this finding is beyond the scope of the study. Note, however, that such a discrepancy could result if the replication error is based on subsampling of simultaneous fermentations for a single nutritional medium, while the order in which the media are tested is randomised.

We continue the example assuming that the use of robust standard errors is appropriate. With a type I error of, say, 10 %, we see from Tables 1-3 that the least significant effect is 1.62, 1.70, or 1.93 times the standard error for ASE, PSE(50), and PSE(45), respectively, resulting in values of 13.21, 15.99, and 20.76, respectively. The effect of ammonium chloride is larger than the least significant effect for ASE, but not so for both versions of PSE. The remaining effects are all smaller than the smallest of the three least significant effects. We conclude that there is some evidence that ammonium chloride affects the enzyme activity. In view of its sign, the effect is not in the direction sought.

3.2 Assessment of some critical factors in the freezing technique for the cryopreservation of precision-cut rat liver slices

Introduction

Precision-cut tissue slices are extensively used in studies on toxicity and metabolism of xenobiotics (Bach *et al.*, 1996; Olinga *et al.*, 1997). The use of slices for this purpose has several advantages as compared to the use of isolated single cells. First, no digestive enzymes are required for slice preparation and, consequently, cell to cell contacts and normal tissue architecture remain intact. Secondly, slices can be used to study species differences, e.g., in drug metabolism (Krumdieck *et al.*, 1980), since they can easily be prepared with the same method from tissue of numerous species. The use of human tissue in this field is of special interest, since it will help overcome difficulties with interspecies extrapolation. The supply of human tissue for *in vitro* research, however, is limited and irregular. Successful cryopreservation and storage of tissue slices, without a marked loss of viability after thawing, would without doubt expand and facilitate experimental possibilities.

The main concern in cryopreservation is the formation of intracellular ice (IIF) both during freezing and thawing. IIF might be the main cause of cellular damage; it can be prevented or reduced using two different approaches. First, slow freezing techniques lead to cellular dehydration. Due to ice crystal formation in the extracellular medium, solutes concentrate in the non-frozen phase. As a reaction to maintain osmotic equilibrium with their surroundings, cells dehydrate and do not freeze intracellularly (Mazur, 1984). Second, rapid freezing techniques, generally applied in the presence of high concentrations of cryoprotectant, result in the formation of amorphous instead of crystalline ice in the cellular fluid (vitrification) (Fahy, 1990).

For the cryopreservation of liver slices, both slow and fast freezing techniques have been described (for a recent overview, see Glöckner *et al.*, 1998). Results, however, have been shown to be highly variable and seem to be dependent on the animal species and the viability parameters used (Ekins, 1996; Ekins *et al.*, 1996; Fisher *et al.*, 1991; Fisher *et al.*; 1993; Wishnies *et al.*, 1991). The identification of important factors that determine cell viability after cryopreservation is difficult due to numerous differences between the various methods published, including the viability parameters determined, the equipment used for freezing and the incubation system for culturing slices.

For cryopreserving rat liver slices, the applicability of a slow freezing method was recently studied, showing that cryopreserved rat liver slices can be used for short term (it is ca. two h) metabolism studies (Maas *et al.*, submitted). Also recently, a fast freezing technique was described (De Kanter and Koster, 1995; De Kanter *et al.*, 1998) reporting cryopreserved rat liver slices to be viable for three h after thawing. The reported time period of cell survival in these studies, although promising, may be too short for prolonged toxicological and metabolism research.

The aim of this study was to determine important factors in the freezing technique in order to find an optimised approach to the cryopreservation of liver slices. For this purpose,

two cryopreservation protocols were compared in a single sub-study. The protocols use either a slow (Maas *et al.*, submitted) or fast freezing approach (De Kanter and Koster, 1995). The sub-study was performed under identical experimental conditions (slices obtained from the same animal cultured in the same incubation system using the same medium), allowing the most direct comparison possible. A number of different parameters were measured to determine which cellular processes were most affected by cryopreservation. The two cryopreservation protocols studied differ not only in in freezing rate, but also in the cryopreservation medium, in the slice thickness, and in the way that the cryoprotectant is added to the slices. Therefore, in a second sub-study, the variables in the two methods were looked at more closely to determine the importance of each variable in maintaining slice viability after cryopreservation. For this purpose, a set of experiments was performed according to a full factorial statistical design, allowing all the variables to be studied simultaneously. The statistical analysis of this set was carried out with the methods of Study 3.1.

Materials and methods

Chemicals

Chemicals were obtained from the following suppliers; Gibco BRL (Paisley, Scotland): Phosphate Buffered Saline (PBS), William's Medium E (WME with Glutamax I) and gentamycin; PAA Laboratories GmbH (Linz, Austria): Foetal Calf Serum (FCS); Sigma Chemical Company (St. Louis, MO, USA): insulin (from bovine pancreas), dimethylsulfoxide (Me_2SO), 1-chloro-2,4-dinitrobenzene (CDNB) and the blood urea nitrogen (BUN) assay kit; Merck (Darmstadt, Germany): D-Glucose; Lamepro b.v. (Raamsdonksveer, the Netherlands): the University of Wisconsin solution (UW); Omnilabo (Breda, the Netherlands): ATP assay kit (Lumac biomass assay kit); Steraloids, Inc., (Wilton, NH, USA): testosterone and hydroxylated metabolites; Pierce (Oud-Beyerland, the Netherlands): the Coomassie Protein kit No.23200; Boehringer (Mannheim, Germany): the lactate dehydrogenase (LDH) kit.

Slice preparation

Male Wistar rats were sacrificed by decapitation. After excision, the liver was cut in lobules that were stored in washing medium (WME containing 10% FCS) on ice. Slices of two different thicknesses were prepared from 8 mm biopsies in cold WME (pregassed with carbogen (95% O_2/5% CO_2) for approx. 30 min), using a Krumdieck tissue slicer. Slices were washed in cold washing medium, and stored in fresh washing medium on ice until use. The time between the sacrifice of the rat and the start of cryopreservation was ca. 2 h.

Culturing

Slices were cultured under carbogen atmosphere in 25 ml Erlenmeyer flasks (1 slice/flask) containing 5 ml culture medium consisting of WME with Glutamax I containing 0.1 µM

insulin, 5% FCS, 50 µg/ml gentamycin and 25 mM D-glucose, pre-gassed with carbogen. Flasks were tightly closed with a rubber stop.

Cryopreservation

The slow freezing technique (Maas *et al.*, submitted) is further referred to as method A and the fast freezing technique (De Kanter and Koster, 1995) as method B and they are described in detail by the aforementioned authors. Table 18 presents a schematic overview. Method A involves computer-controlled slow freezing using a SyLab Icecube 1610 Computer Freezer. Slices were transferred to ice-cold UW solution containing 5% Me_2SO and were stored on ice for approximately 10 min. Slices were then transferred to cryovials (Costar, Cambridge, MA, USA) (5 slices/vial) containing 0.5 ml UW/10% Me_2SO and slowly frozen at 0.5 ^oC/min to -50 ^oC followed by a 1 ^oC/min freezing rate to -80 ^oC, after which the vials were submerged in liquid nitrogen. Injecting extra nitrogen into the freezing chamber prevented a large temperature rise in the sample vial due to the release of heat during crystallisation. Total pre incubation time at 0-2 ^oC was ca. 30 min.

With method B, slices were pre incubated before freezing with 5 ml WME containing 12% Me_2SO in an 25 ml Erlenmeyer (max. 5 slices/flask, under carbogen atmosphere) in a shaking waterbath (at ca. 100 rpm., at 2 ^oC for 30 min). Slices were then transferred to 2 ml cryovials (4 or 5 slices/vial) together with 1 ml of the medium used for pre-incubation and frozen by direct immersing into liquid nitrogen.

Sub-study I

In this sub-study, method A and B were compared under standardised experimental conditions. A number of viability parameters was studied to determine which parameters were most sensitive to cryopreservation. The viability parameters studied were adenosine triphosphate (ATP) and potassium (K^+) levels, histomorphology, glutathione (GSH) levels leakage of lactate dehydrogenase (LDH), total glutathione-S-transferase (GST) activity, urea formation and the metabolism of 1-chloro-2,4-dinitrobenzene (CDNB) and of testosterone.

Table 18: Schematic overview of the two cryopreservation methods used in Sub-study I

	method A (Maas *et al.*, submitted)	method B (De Kanter *et al.*, 1995)
Wet weight slices (mg)	16 - 18	19 - 21
Slice thickness (µm)[a]	± 200-250	± 250-300
Addition of cryoprotectant	stepwise addition	immediate addition
Freezing rate	0.5 ^oC/min	± 250 ^oC/min
Cryopreservation medium	UW	WME
Cryoprotectant concentration	10% Me_2SO	12% Me_2SO

[a]In a calibration study (results not shown), series of slices were cut at different settings of the Krumdieck tissue slicer. Wet weight was correlated to slice thickness, which was determined morphometrically. In further studies, wet weight of the slices was used to set slice thickness.

Experimental design

The experimental design of Sub-study I is presented in Table 19. Three separate experiments were performed. All parameters were determined at three time points after slice preparation and thawing and incubation (0, 4 and 8 h) in three slices. Leakage of LDH was determined during two successive periods of four h after isolation or thawing of the slices (three per group).

Viability parameters

Histomorphology. One half of a slice was fixed in 70% ethanol at 4 ^{o}C for at least 24 h. After dehydration, slices were vertically embedded in paraffin and cross sections (5 µm thickness) were stained with haematoxylin and eosin. Histomorphological quality of the slices was scored by estimating the percentage of viable cells in the slice cross section by two investigators. For this purpose, both nuclear appearance and cytoplasmatic staining were taken into account and were compared to cell appearance in fixed samples from fresh liver. Slice thickness was evaluated by counting the number of cell layers.

ATP content. One half of a slice was quickly rinsed in physiological saline and immersed in cold 70% ethanol in potassium free HPLC water (Baker, Deventer, the Netherlands), containing 2 mM EDTA (pH 10.9). The slice was homogenised using a Branford sonifier (50% duty cycle, 5 s). ATP was determined using the luciferin/luciferase assay on a Lumac Biocounter M500 (Singh *et al.*, 1996).

Potassium content. In the slice homogenate, potassium content was measured using a Beckman 2 electrolyte analyser (Beckman Instruments-Netherlands division BV, Mijdrecht, the Netherlands).

GST activity. GST activity in the slice homogenate was determined on a Varian Cary IE spectrophotometer using 1 mM CDNB and 1 mM GSH as substrates. S-(2,4-dinitrophenyl)glutathione (DNPSG) formation was measured for three min at 25 ^{o}C (detection wavelength: 340 nm).

Metabolism of testosterone. To determine the metabolism of testosterone, slices were incubated for 60 min. with 250 µM testosterone in 12-well plates (Costar, Cambridge, MA, USA; 2 ml/well) on a gyratory shaker in a humidified incubator at 37 ^{o}C, 40% O_2/5% CO_2 at ca. 80 rpm. Hydroxylated metabolites of testosterone were determined by HPLC on a Hypersil ODS column (Chrompack, the Netherlands) according to Van 't Klooster *et al.* (1993).

Metabolism of CDNB. For determining CDNB metabolism, slices were incubated for 15 min with 25 µM CDNB in 12-well plates (2 ml/well) on a gyratory shaker in a humidified incubator at 37 ^{o}C, 40% O_2/5% CO_2 at ca. 80 rpm. DNPSG was quantified by HPLC using a Zorbax ODS column according to Van Iersel *et al.* (1996).

GSH content. GSH was determined using a Shimadzu RF1501 spectrofluorimeter after reacting

Table 19. Experimental set-up of Sub-study I[a]

		Sampling/incubation time (h)		
Experiment	Viability parameter:	0	4	8
I	GSH, GST activity,	0	4	8
	LDH activity	-	0-4	4-8
	Urea synthesis	0-3	4-7	8-11
II	ATP, K^+ and histomorphology	0	4	8
	testosterone hydroxylation	0-1	4-5	8-9
III	CDNB metabolism	0-0.15	4-4.15	8-8.15
	histomorphology	0	4	8

[a]This sub-study consisted of three independent experiments.

with o-phtalaldehyde (OPT) according to Hissin and Hilf (1976).

Lactate dehydrogenase activity. LDH activity in the medium and the slice homogenate was measured with a BM/Hitachi 911 using a commercially available kit (Boehringer, Mannheim, Germany).

Total protein. To an aliquot of the slice homogenate, 2 M NaOH (25% v/v) was added to dissolve the protein. Hereafter, the protein solution was diluted (1:5 minimally) with Phosphate Buffered Saline. Subsequently, protein content was measured by a Cobas-Bio centrifugal spectrophotometer with a Coomassie Protein kit using Bovine Serum Albumine (BSA) as a standard.

Statistical analysis.

The data were analysed by pairwise comparisons of mean values for the methods A and B, respectively. Comparisons were carried out separately for fresh and cryopreserved slices at each of the time points. Statistical significance was assessed by *t* tests using log-transformed data. Variances of the groups of fresh slices were pooled to obtain a standard error for the t testing. The same procedure was carried out to obtain standard errors for the cryopreserved slices.

Sub-study II

The aim of the second set of experiments was to specify the variables between method A and B that could explain differences in slice viability as found in Sub-study I.

Experimental design

The two cryopreservation methods compared in Sub-study I not only differ in freezing rate but also in the cryopreservation medium used, slice thickness and the way of addition of the cryoprotectant (see Table 18). A set of two experiments was carried out according to a full two-level factorial design. In this design, each of the variables that make up the difference between the two cryopreservation methods was studied at all combinations of levels of the other variables as illustrated in Table 20. The design contained four factors representing the main variables between method A and B. Furthermore, the factor ‘experiment’ was introduced to assess the reproducibility of the effects.

Viability parameters

The viability parameters used in Sub-study II were ATP, K^+, histomorphology and GST activity. They were selected from Sub-study I in view of their sensitivity and the ability to distinguish between the two cryopreservation methods. All viability parameters were assessed in the same slice. They were measured as described for Sub-study I.

Statistical analysis

Data obtained directly after thawing or after an additional culture period of 4 h were analysed separately. Statistical analyses were carried out on the natural logarithm of the biochemical parameters and the square root of thickness (number of cell layers). For

histology scores, the empirical logistic transformation was used, defined as transformed score = log [(score + 0.5)/ (100 - score + 0.5)] (Cox, 1970). Statistical analyses were based on the mean transformed data from two slices for each parameter (ATP, K^+, histomorphology and GST activity). The outcome of each of the combinations (illustrated in Table 20) depends on effects that are caused by the four handling factors, the factor 'experiment' and the random variation. From the 32 values (mean of two slices) for each parameter (16 possible combinations in two experiments), a total of 31 effects can be calculated (namely, 5 main effects, 10 two-factor interactions, 10 three-factor interactions, 5 four-factor interactions, and 1 five-factor interaction). To illustrate this, the main effect of the medium is calculated as the mean of all observations made with the UW medium minus the mean of the observations with the WME medium. Other main effects are calculated analogously. Two-factor interactions occur in those cases were the observed effect cannot be explained solely by the main effects. The interaction between medium and freezing, for example, is calculated as the difference in the main effect of medium at slow freezing with the main effect of medium at fast freezing. This value is divided 2, causing all interactions to have the same standard error as the main effects. The remaining interactions are calculated in a way analogous to the calculation of the two-factor interactions. The design and the calculation of the effects are described by Box *et al.* (1978). Statistical significance

Table 20. Experimental design of Sub-study II[a]

Combination	Factor (variable)			
	Medium	Cryoprotectant addition	Freezing speed	Slice thickness[b]
1	UW	stepwise	fast	thick
2	UW	immediate	fast	thick
3	UW	stepwise	fast	thin
4	UW	immediate	fast	thin
5	WME	stepwise	fast	thick
6	WME	immediate	fast	thick
7	WME	stepwise	fast	thin
8	WME	immediate	fast	thin
9	UW	stepwise	slow	thick
10	UW	immediate	slow	thick
11	UW	stepwise	slow	thin
12	UW	immediate	slow	thin
13	WME	stepwise	slow	thick
14	WME	immediate	slow	thick
15	WME	stepwise	slow	thin
16	WME	immediate	slow	thin

[a]This sub-study consisted of two experiments. In both experiments, four slices were treated according to each combination. Directly after thawing and after four h of additional culture, two slices were sampled for the determination of slice viability. Besides the four factors that represent the main variables between method A and B, the factor 'experiment' was introduced (see materials and methods). [b]The actual difference in thickness between thin and thick slices was determined at the end of the experiments by counting the number of cell layers in the slides that were made for histomorphological examination. The average number of cell layers in the slices in two experiments was 13.4 for thick slices and 10.7 for thin slices.

of the 31 effects on each viability parameter was determined by calculating the standard error of the effects with the PSE(45) method as described in Study 3.1, and using this standard error to judge the effects. Effects were considered significant ($P < 0.05$) if they are larger then 2.35 times their standard error (see Table 15).

Results

Sub-study I

Histomorphological examination

Directly after preparation of the slices, the average percentage of viable cells in the slice was scored 85 ± 5%. After 4 and 8 h in culture, this percentage was slightly decreased to 74 ± 4 and 69 ± 5%, respectively. Immediately after thawing, the histomorphogical views of slices frozen according to method A and B closely resembled each other. Compared to freshly prepared slices, a slight increase in the number of pyknotic nuclei was observed and cytoplasmic staining was slightly more eosinophilic. The percentage of viable cells per slice was 63 ± 12%. Differences between slices frozen according to method A and B were more pronounced after an additional four h of culturing, where the percentage of viable cells per slice ranged from 0 to 5% and 5 to 50% for method A and B respectively. After eight h of culturing, the percentage of viable cells frozen according to method B was decreased, ranging from 5 to 25%. In some slices, viable cells appeared in small groups in the venal area and at the edges of the slice. Many dead cells had fragmented nuclei. In slices frozen according to method A, only a few viable cells were observed randomly distributed over the slice. Affected cells showed mostly pyknotic nuclei.

Data obtained from fresh and cryopreserved rat liver slices are presented in Fig. 4. For fresh slices, slice thickness was the only difference between the two methods. Data from cryopreserved slices will be discussed below.

ATP, K^+ and GSH contents

After cryopreserved slices were thawed and subsequent cultured for 4 and 8 h, ATP and K^+ levels in slices cryopreserved according to method B were significantly higher compared to levels in slices cryopreserved according to method A. No difference was observed in GSH content. K^+ content in slices cryopreserved according to method B was ca. 20% of the K^+ content measured in fresh slices, both after 4 and 8 h in culture.

LDH leakage

LDH leakage from slices cryopreserved according to method A was significantly higher compared to leakage from slices frozen according to method B. Cryopreservation according to method A and additional culturing for four h, resulted in 90% leakage of total LDH. For

slices frozen according to method B, leakage of LDH increased from 60% to 80% of total LDH after four and eight h in culture respectively.

GST activity

Total GST activity in the homogenate of slices frozen according to method A and B significantly differed at all time points. In slices cryopreserved according to method A, GST activity declined to 15% of values in fresh slices after 4 h and remained at the same level after 8 h in culture. In slices frozen according to method B, respectively 67% (after 4 h) and 58% (after 8 h) of total GST activity remained.

Metabolism of CDNB

In cryopreserved slices, CDNB metabolism decreased with time with only limited formation of DNPSG left after eight h. DNPSG formation in slices cryopreserved according to method B was significantly higher from the formation in slices frozen according to method A, only after four h of culture.

Metabolism of testosterone

No significant difference in the formation of all hydroxylated metabolites (OHT) determined in this sub-study (2α-, 6ß-, and 16α-OHT) was found between fresh and cryopreserved slices at any time point (both in slices cryopreserved according to method A and B). Fig. 4 shows just the formation of 2α-OHT.

Urea synthesis

Following cryopreservation, urea synthesis in slices cryopreserved according to method B was significantly higher than in slices frozen according to method A at all time points. After 8 h in culture, urea synthesis in slices frozen according to method A and B was 17% and 44% of fresh values, respectively.

Sub-study II

In this sub-study, variables between the two cryopreservation methods (that is freezing rate, cryopreservation medium, slice thickness and the mode of cryoprotectant addition) were studied more closely in order to explain the differences in slice viability found in Sub-study I.

Determination of critical factors (significant effects) directly after thawing

The main effects of all five variables on the selected viability parameters directly after thawing are presented in Fig. 5, left side. In slices cryopreserved according to method A, relatively high potassium levels were measured, resulting in a significant effect of the factor 'medium' ('wme/uw' in Fig. 5) on the parameter K^+. Furthermore, based on the same

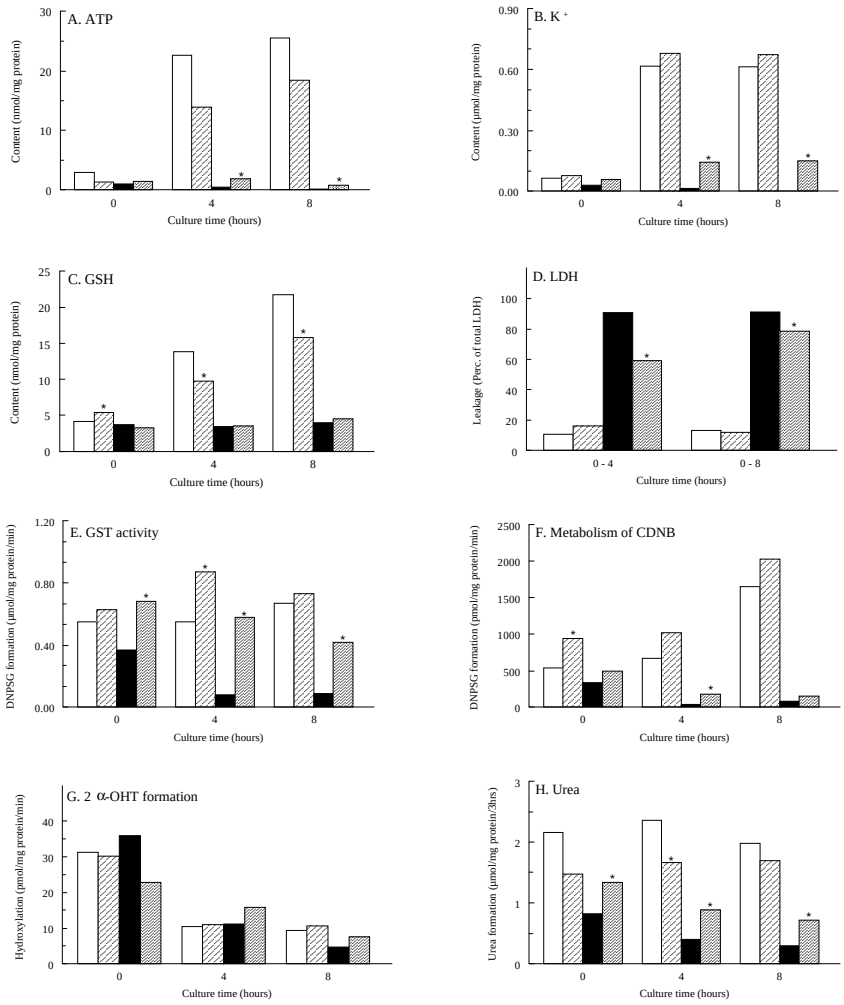


Fig. 4. Sub-study I: viability parameters determined in fresh and cryopreserved rat liver slices, directly after isolation or thawing and after 4 and 8 h of additional culture. No dash: method A, fresh; light dash: method B, fresh; black: method A: cryopreserved; heavy dash: method B, cryopreserved. Statistical analysis was carried out separately for fresh and cryopreserved slices at each of the time points. * Significantly different, $P \leq 0.05$.

parameter, slice thickness was determined to be decisive, with thick slices containing seen between the two experiments ('II / I' in Fig. 5) on the parameters ATP and GST slightly higher potassium levels than thin slices. Finally, a significant difference was also activity.

Determination of critical factors (significant effects) after 4 h of culture

The main effects of all five variables on the selected viability parameters after 4 h of culture are presented in Fig. 5, right side. The choice of the cryopreservation medium significantly affected the outcome of the viability parameters histomorphology, GST activity and K^+. Slices cryopreserved in WME showed better histomorphology and higher GST activity and K^+-levels, after four h of culturing. Also the freezing rate significantly affected the outcome of the same parameters. This finding was consistent in both experiments of Sub-study II.

For ATP, the observed effect could not solely be explained by the effect of the main variables alone. Indeed, a significant interaction was observed between freezing rate and medium. This interaction between freezing speed and medium was also found on the parameters histomorphology and GST activity. For K^+ content, a significant (three factor) interaction between freezing, medium and cryoprotectant addition was observed. These interactions are presented in Fig. 6. The use of UW instead of WME caused a decrease in all parameter levels in fast frozen slices.

For all viability parameters determined, highest levels were measured in slices that had been fast frozen in WME, independent of slice thickness. The average K^+ level was 0.21 ± 0.03 µmol/mg protein which was ca. 30% of the value measured in fresh slices. ATP levels in Sub-study II showed a large variation (as illustrated in Fig. 5), ranging from 2% to 14% in the first experiment and from 2% to 25% in the second experiment (both compared to fresh slices). Considering histomorphology, the average percentage of viable cells was 10 ± 9%. The highest percentage of viable cells scored in Sub-study II was 35%. The average GST activity was 33 ± 15% and 53 ± 18% (both compared to fresh slices) in experiments I and II, respectively.

Discussion

The aim of this study was to evaluate important factors in the freezing process in an effort to find an optimised approach for the cryopreservation of precision-cut liver slices. To meet this purpose, a comparative sub-study was carried out using two freezing techniques for the cryopreservation of rat liver slices (De Kanter and Koster, 1995, Maas *et al.*, submitted) to establish any differences in tissue viability for a number of endpoints. Subsequently, a second sub-study was performed to identify the most important variables between the two cryopreservation methods that could explain the differences in slice viability after cryopreservation, as observed in the first sub-study. The second sub-study was performed according to a multifactorial experimental design, that has been shown to be an effective way to study main effects and interactions of several variables in toxicological research with different end-points (Groten *et al.*, 1991; Section 4.1).

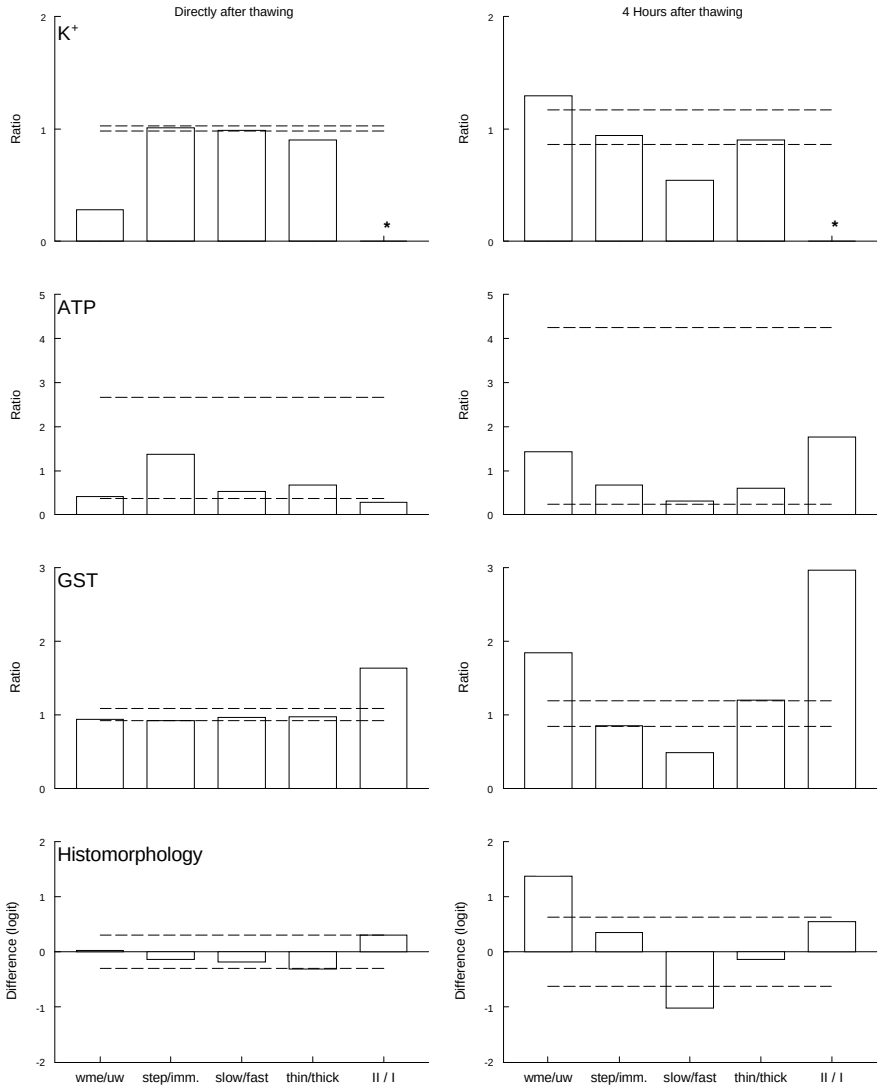


Fig. 5. Sub-study II: main effects of experimental factors on viability parameters directly after thawing (left side) and after 4 h of additional culturing (right side). Boundaries between which the effects are nonsignificant ($\alpha = 0.05$) are given with dashed horizontal lines. The Y axis represents the ratio between geometric means taken at both levels of a factor. For histomorphology, effects are represented as differences on a logit scale. step/imm: stepwise addition/immediate addition of cryoprotectant. *: K^+ content was determined in experiment II only.

The comparative sub-study between method A (Maas *et al.*, submitted) and B (De Kanter and Koster, 1995) was performed under identical conditions. Directly after thawing, differences between the two methods were limited to urea synthesis and GST activity, with method B showing significantly higher levels for both parameters. After 4 and 8 h of culture, viability of slices cryopreserved according to method B was significantly higher than that of slices frozen according to method A, based on all parameters with the exception of GSH content and testosterone hydroxylation.

A clear difference between the viability parameters was observed with respect to their sensitivity to cryopreservation. Urea synthesis and GST activity in the slice homogenate were relatively well maintained with method B, while the metabolism of testosterone was maintained by both methods. This latter finding is consistent with the report that membrane bound enzymes, such as the cytochrome P450 complex, are resistant to freezing stress (Ekins, 1996). However, a more pronounced loss of viability was apparent when ATP and K^+ content and histomorphology were considered.

In the second sub-study we tried to specify the important variables that could explain the differences between the two cryopreservation techniques that were found in the first sub-study. Directly after thawing, a significant medium effect was observed on the parameter K^+. Furthermore, a significant experiment effect was detected on GST activity and ATP content. These effects, however, were considered of less importance since they could be explained by the high concentration of potassium in the UW solution (that was not complete removed from the slice during washing), or biological variation.

After 4 h in culture, two variables were detected that could explain to a large extent the differences between the two cryopreservation methods, namely freezing speed and the cryopreservation medium. Viability of slices that were rapidly frozen using WME as cryopreservation medium was highest compared to all other groups. This effect was observed on all parameters except ATP.

With slow freezing of rat liver tissue, it was recently shown that in the presence of Me_2SO, cellular dehydration continued below -20 ^oC until maximum dehydration was reached (Smith *et al.*, 1998). In the present study, the Me_2SO concentration was in the same range as that used by Smith *et al*. (1998) and slow freezing was continued until -50 ^oC. We would therefore expect almost maximum cellular dehydration to occur, with minimal IFF. However, long term exposure (caused the low freezing rate) of cellular membranes to concentrated intracellular solutes (resulting from cellular dehydration), may have caused cellular damage. With the fast freezing approach used here, we expect intracellular ice to form, since the freezing rate does not allow cellular dehydration and both freezing rate and cryoprotectant concentration are too low to allow for vitrification to occur. However, our results indicates that fast freezing is still favored over slow freezing for the cryopreservation of rat liver slices. We can speculate that the ice crystals formed at least in some of the cells, are small, and therefore cellular damage is limited. It can be hypothesised that more rapid freezing and thawing (ice crystals tend to be small at high cooling rates; Mazur, 1984) might contribute to higher survival rates. Only recently, it was demonstrated that a significant increase in liver slice viability could be achieved when aluminium plates instead of cryovials were used for freezing. Due to the much higher thermal conductivity of aluminium, higher freezing rates could be reached (Day *et al.* 1999). Preliminary data using small stainless steel grids and higher cryoprotectant concentrations, do confirm these observations (Maas *et al.*, De Graaf *et al.*, unpublished results).

In addition to the freezing speed, the cryopreservation medium was found to be an important factor determining the success of cryopreservation; the use of WME resulted in higher viability than the use of UW, which is successfully used in cold storage of tissue. This finding was unexpected, since we believed the effect of the UW solution to be beneficial by it preventing cold induced cell swelling (Southard and Belzer, 1993) leaving less unbound water in the cells to crystallise.

No effect of cryoprotectant addition or slice thickness was observed in the present study. Due to the much lower cell membrane permeability of Me_2SO than of water, the rapid introduction of cryoprotectant into the medium can cause an osmotic shock leading to cell damage (Levin and Miller, 1981). In the present study, no significant beneficial effect of

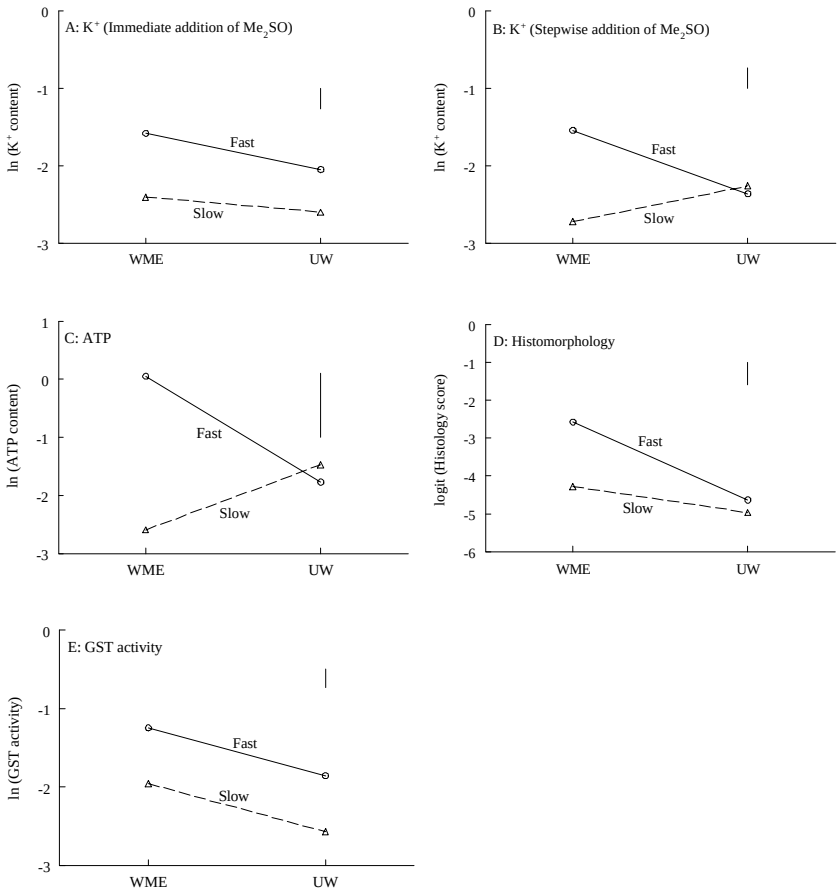


Fig. 6. Sub-study II: joint effects (interactions) of experimental factors 4 h after thawing and subsequent culturing. A: K^+, with immediate addition of the cryoprotectant; B: K^+, with stepwise addition of the cryoprotectant; C: ATP content; D: histomorphology; E: GST activity. Vertical lines: least significant difference (α=0.05) between points in a graph.

stepwise addition of the cryoprotectant on slice viability was observed. Most likely, the cryoprotectant concentration used was low enough to prevent osmotic damage.

With respect to slice thickness, this study showed no major effect of slice thickness on viability of the slices after cryopreservation. The actual difference between thin and thick slices was statistically significant (average difference of ca. three cell layers). Theoretically, slice thickness could be an important factor in the success of cryopreservation. Thermal gradients may occur in thick slices, leading to differences in freezing rate at the edges and in the middle of the slice. Furthermore, penetration of cryoprotectant in thick slices could be insufficient. Indeed, in some slices that had been fast frozen in WME, viable cells were observed in groups located around the venal area or at the edges of the slice. This indicates that cryoprotectant penetration or freezing rates may have been insufficient in cells located deeper in the tissue, and that the use of thin slices is favourable. The reason slice thickness was not found to be a determining factor is most likely that the difference in slice thickness was too small to cause a clear effect on viability.

In summary, this study shows that hydroxylation of testosterone, urea synthesis and GST activity were relatively insensitive to cryopreservation stress. ATP and K^+ contents and histomorphology were more clearly affected and can therefore be considered more sensitive parameters for evaluating slice viability after cryopreservation. Freezing speed and the cryopreservation medium were demonstrated to be major determinants of the residual viability of liver slices after cryopreservation and subsequent culturing.

3.3 Choosing appropriate two-level experiments to detect location and dispersion effects

Introduction

The current interest in reduction of product variation gave rise to procedures detecting the influence of experimental factors on the random variation, or dispersion, of the response of interest. These procedures may be applied as early as the screening phase of the experimental investigation of products or processes (Box and Meyer, 1986; Bergman and Hynén, 1997; Engel and Huele, 1996; Nelder and Lee, 1991; Lee and Nelder, 1998; Nelder and Lee, 1998). There are roughly two approaches. Firstly, we have procedures studying dispersion effects separately from location effects. Both the first and the latest of these procedures are based on F tests of squared contrasts that are orthogonal to the location effects. These will be designated 'VR'. The location effects are fitted either on the whole dataset (Box and Meyer, 1986), or on the halves of the dataset that correspond with the dispersion effect to be tested (Bergman and Hynén, 1997). Wang (1989), and Fuller and Bisgaard (1995) give alternative non-iterative methods. These are based on score statistics, and absolute residuals, respectively.

An alternative approach to detect dispersion effects uses a joint modelling of mean and dispersion, as first proposed by Aitkin (1987) for normal errors. The model fitting involves a seesaw algorithm to alternate between the fit of a location model and a dispersion model. The alternation ends when the normal likelihood has converged to a stable value. The location model has weights obtained from the dispersion model, while the dispersion model is based on residuals from the location model. Detection may proceed through likelihood ratio tests for various models. For this reason, we will designate this approach as 'LR'.

All dispersion-effects detection methods presuppose a valid location model. If, however, the location model has to be found empirically, then misspecifications may result. Pan (1999) studied the impact of unidentified location effects in dispersion-effects detection with VR methods. He demonstrated that unidentified location effects could have a very severe impact on the detection of dispersion effects. The pattern of the impact is unpredictable if the true location model is unknown.

The present study bears on the use of two-level fractional and full factorial experiments to detect both location and dispersion effects. Given the possibility of one or more active dispersion effects, given several methods for their detection, and given the severe impact of misspecifications in the location model, it addresses the following practical issues. Firstly, in the presence of active dispersion effects, the usual estimates of the location effects are no longer independent. The popular method of normal or half-normal plotting of the effects (Daniel, 1959, 1976), and more formal methods based on robust scale estimates as the Pseudo Standard Error (PSE; Lenth, 1989; Haaland and O'Connell, 1995; Section 3.1) presuppose independence. While the graphical methods are hard to approach algorithmically, we can do so with the PSE-based tests. Do these tests still give reasonable results?

As a second issue, we study to what extent the dispersion-effects detection can be improved if we can make assumptions on the order of the location model. More

specifically, we will consider experiments with only main effects, and experiments with some additional interactions.

Thirdly, LR methods to detect dispersion effects may not suffer from the problems with undetected location effects in the same way as the VR methods do. We will study the performances of a VR method, and one specific LR method likely to be useful in screening experiments, under various effect configurations to see if there is a clear winner.

Further, as there is aliasing in fractional designs, we will study the impact of the resolution of a design on its possibilities to detect dispersion effects. Note that Box and Meyer (1986) as well as Bergman and Hynén (1997) use resolution III designs as examples.

Finally, dispersion-effects detection in replicated designs does not suffer from the problems with unidentified location effects. We will compare performance measures in these designs with those of unreplicated designs that have equal size, but better resolution for the location effects.

This study addresses the above practical issues with the VR method of Bergman and Hynén (1997) and an LR method with a fitting algorithm proposed by Lee and Nelder (1998) as vehicles. The methods are subjected to theoretical considerations. They are also used in simulations of location and dispersion-effects detection in two-level experiments under a wide range of specifications. The organisation of this study is as follows. We start with the specification of the experimental models considered here, and review the detection strategy and the methods of detection to be investigated. Subsequently, we demonstrate the detection procedures using a cryopreservation experiment from Study 3.2. Then, we motivate the design of an extensive simulation study using replicated as well as unreplicated designs. The results are given next. This study is concluded by a discussion of the findings.

Detection of active effects

We consider results from two-level experiments that obey the relation

$$y_i = \boldsymbol{x}_i'\beta + e_i$$

with i indexing the run numbers, y denoting the response, $\boldsymbol{x}$ a p-dimensional vector containing explanatory variables varied between the runs, β a p-dimensional vector with the corresponding true effects, and e a random error.

The e_i are postulated to be i.i.d. normally distributed with expectations zero. Their variances are denoted σ^2_i. These are supposedly log-linearly related to the experimental variables according to

$$\log \sigma^2_i = \boldsymbol{u}_i'\gamma$$

The experimental variables thought to influence the variance are in the vector $\boldsymbol{u}$; the corresponding true effects are in the vector γ.

Detection of location effects

To detect dispersion effects with either an LR or a VR method, we need some location model to start with. We find roughly two approaches. The first one uses some maximal model for the location effects, for example by including all main effects. For unreplicated experiments, Lee and Nelder (1998) use this approach as a basis to search for a dispersion model. Note that we can also use a maximal location model in conjunction with a VR method.

From the way in which screening experiments are usually employed, a maximal model for location effects is not likely to be useful for subsequent dispersion-effects detection. In theory, one could define a maximal model containing just the main effects, or a model containing main effects and interactions between two factors. However, the first type of model is hopefully adequate only for resolution III designs. As these designs typically have many main effects, few degrees of freedom are left for modelling the error. The second type of maximal model could be used for designs with a resolution higher than IV. There need not be many contrasts involving higher order interactions. Finally, resolution IV designs are typically employed if one expects a few interactions. Thus, for these designs, a maximal model with just the main effects is not adequate, while a model containing all interactions contra-indicates the use of a resolution IV design.

An alternative for using a maximal model for location effects is to construct a model from effects appearing active from a halfnormal plot (Daniel, 1959) or a full normal plot (Daniel, 1976). Indeed, this was proposed by Box and Meyer (1986), and Bergman and Hynen (1997). Note that this detection method can also be used in conjunction with subsequent joint modelling of mean and dispersion.

Due to its informal nature, we cannot use the judging of halfnormal plots as an element of assessing the detection of both location and dispersion effects. Instead, we use effect tests with a robust estimator of the standard error. Various estimators have been developed, see Haaland and O'Connell (1995) for an overview. We use the Pseudo Standard Error (PSE) of Lenth (1989), as modified by Haaland and O'Connell (1995). It is calculated as PSE = a_{PSE} . median{$|b_j|$: $|b_j| \leq 3.707$. median{$|b_j|$}}, where a_{PSE} is a simulated adaptive constant, and the b_j are the estimated standardised contrasts. The values used for a_{PSE}, and the critical values for effect testing are given in Table 14; see Study 3.1 for a more detailed account of PSE.

Detection of dispersion effects

Variance Ratio method. In this study, we use the Variance Ratio method as given by Bergman and Hynén (1997). For each of the experimental factors, a test statistic F_{VRi} is calculated as $F_{VRi} = \Sigma_j z^2_{j\,|\,i+} / \Sigma_j z^2_{j\,|\,i-}$. Here, $z_{j\,|\,i+}$ and $z_{j\,|\,i-}$ are contrasts in the halves of the data with factor i on its + and – level, respectively. The $z_{j\,|\,i+}$ are mutually orthogonal, and orthogonal to the location effects fitted in the relevant half of the data; the $z_{j\,|\,i-}$ are defined similarly. Bergman and Hynen (1997) note that this is equivalent with taking the residual sum of squares after fitting a location model on each of the halves of the dataset separately. The values of F_{VRi} are compared with critical values of an $F_{[mi;mi]}$ distribution, where $mi = \Sigma_j 1$. This is a two-sided test.

For replicated experiments, we calculate $F_{VRi} = \Sigma_j s^2_{j\,|\,i+} / \Sigma_j s^2_{j\,|\,i-}$, with $j \in \{1\ldots \tfrac{1}{2} n\}$ indexing the treatments in both halves of the dataset, n denoting the number of distinct

treatments and s^2 denoting the replicate variances. We compare F_{VRi} with critical values of an $F_{[q,q]}$ distribution, with $q = ½n(r\text{-}1)$, r denoting the number of replicates of the design.

If the dispersion effect under consideration is the only one that is active, or if there are no active dispersion effects at all, then the contrasts $z_{j\,|\,i+}$, and $z_{j\,|\,i-}$, are indeed orthogonal. However, when there are dispersion effects within the halves of the dataset, the contrasts are correlated. In particular, for contrasts z_m and z_l, $m, l \neq i$,

$$\text{cov}\,(z_{m\,|\,i+}\,,\, z_{l\,|\,i+}) \;=\; \sum_{j|i+,m=l} \sigma_j^2 \;-\; \sum_{j|i+,m\neq l} \sigma_j^2$$

which is non-zero when the generalised interaction of columns m and l corresponds with an active dispersion effect. In general, there are $½(2^{k-p-1}\text{-}2)$ pairs of contrasts that have a particular column in a half of the dataset as their generalised interaction. If just one column of each of these pairs is involved in the dispersion F test, then independence of the pairs is not an issue. If both columns of a pair are involved, then F_{VRi} does not follow an F distribution.

The extension of the VR method to replicated 2^{k-p} experiments suffers from the same kind of problems as the original method. Indeed, when there is a dispersion effect in a factor $s \neq i$. it may be shown that F_{VRi} follows a ratio of sums of a χ^2 and a Γ distribution, which collapses to an F distribution only if there is no dispersion effect of factor s. We conclude that, generally, F_{VRi} does not follow an F distribution if there is at least one active dispersion effect.

A further possible drawback of the Bergman and Hynen VR test is their sensitivity in resolution III designs to the function linking σ^2_i to a model linear in its parameters. To illustrate this, consider a design with C=A*B. Suppose that A and B have active dispersion effects, while C is not active. Table 21 shows the expectations of F_{VRi} for the factors A, B, and C, respectively, under a correct location model, and under either a linear or a log-linear model for the variance. From this table, it is clear that the VR test for factor C does not work under a log-linear model. Nelder and Lee (1998) implicitly pointed this out when they observed that Bergman and Hynén (1997) assumed an identity link. Indeed, this may explain the dispersion effect in column 2 (F_{VR2} = 15.93) of the welding strength example from the latter authors, because this column is the generalised interaction of columns 13 (F_{VR13} = 20.96) and 15 (F_{VR15} = 21.72). If the true model for the variance were log-linear, and if the realisations of F_{VR13} and F_{VR15} were close to their expected values, then the expected value of F_{VR2} in absence of a real dispersion effect would be about 10.69. This is close to the realisation of F_{VR2}.

Likelihood Ratio method. In the present study, we performed LR tests to compare a null

Table 21. Expected values of F_{vri} under a linear and a log-linear model for the variance, assuming a correct location model

factor	model	
	$\sigma^2_i = \gamma_0 + \gamma_A A + \gamma_B B$	$\ln \sigma^2_i = \kappa_0 + \kappa_A A + \kappa_B B$
A	$(\gamma_0 + \gamma_A A)/ (\gamma_0 - \gamma_A A)$	$\varphi_A = \exp\{2\kappa_A\}$
B	$(\gamma_0 + \gamma_B B)/ (\gamma_0 - \gamma_B B)$	$\varphi_B = \exp\{2\kappa_B\}$
C = A*B	1	$1 + (\varphi_A - 1)(\varphi_B - 1)/ (\varphi_A + \varphi_B)$

dispersion model with a model containing a single dispersion main effect, for each of the main effects under investigation. The test statistics, denoted with X^2_{LRi}, are calculated as follows.

1. Obtain –2 times the residual log-likelihood L_c of a model with just the active location effects from

$$-2\log L_c = \sum_{j=1}^{n} \frac{(y_j - x_j'\hat{\beta})^2}{\hat{\sigma}^2} + n\log(2\pi\hat{\sigma}^2) + \log|(X'X)/2\pi\hat{\sigma}^2|$$

   where X is the design matrix for the detected location effects.
2. Obtain –2 times the residual log-likelihood L_{c+i} of a joint model for mean and dispersion, with the detected location effects and a dispersion effect for factor i by using the following seesaw algorithm.

a. Obtain the OLS estimates of β; set $s^2_j = 1$.

b. Calculate the squared leverage-corrected residuals $r^2_j = (y_j - x'_j\hat{\beta})^2 / (1 - h_j)$, where the leverage h_j is defined as the jth diagonal element of the hat matrix $H = X(X'V^{-1}X)^{-1}X'V^{-1}$, with V a diagonal matrix with entries s^2_j.

c. Fit a generalised linear model to the r^2_j with a $\sigma^2_j \chi^2_{[1]}$ error distribution, and $\log E(r^2_j) = \log \sigma^2_j = \lambda' z_j$, with $z_j' = (1\ i_j)$, i_j denoting the jth setting of factor i; use $(1\text{-}h_j)$ as weights.

d. Calculate $s^2_j = \exp(\hat{\lambda}' z_i)$

e. Fit a linear location model using weights $1/s^2_j$

f. Repeat (b) to (e) until convergence of the criterion Q_{c+i}, with

$$-2Q_{c+i} = \sum_{j=1}^{n} \left\{ \frac{(y_j - x_j'\beta)^2}{(1-h_j)s_j^2} + \hat{\lambda}' z_j \right\} + n\log(2\pi)$$

g. Evaluate the log-residual likelihood, L_{c+i}, with

$$-2\log L_{c+i} = \sum_{j=1}^{n} \left\{ \frac{(y_j - x_j'\beta)^2}{s_j^2} + \hat{\lambda}' z_j \right\} + n\log(2\pi) + \log|(X'V^{-1}X)/2\pi|$$

3. Obtain $X^2_{LRi} = 2\log(L_{c+i} / L_c)$, and judge the activity of factor i by comparing this to percentage points of the $\chi^2_{[1]}$ distribution.

The seesaw algorithm was first proposed by Aitkin (1987). This author used the normal likelihood as fitting criterion. The use of the residual likelihood was proposed by Verbyla (1993); see also Nelder and Lee (1998). Since the convergence of the seesaw algorithm is generally slow, Lee and Nelder (1998) proposed initial minimisation of $-2Q_{c+i}$, with a subsequent evaluation of $-2\log L_{c+i}$. For normal errors, the estimating equations for λ that follow from the REML score function can be shown to be a weighted version of the equations that result from minimisation of $-2Q_{c+i}$ with weights $(1\text{-}h_j)$ (Huele *et al.*, submitted)

LR method for replicated designs. LR tests for replicated designs use the sample variances as response variable and a standard generalised linear model with a $\sigma^2_j \chi^2_{[1]}$ error distribution (McCullagh and Nelder, 1989). The tests are comparisons between a null

dispersion model with a model containing a single dispersion main effect, for each of the main effects under investigation.

ANOVA on logged sample variances. Fitting a linear model with main effects only on the logged sample variances permits calculation of $LS_i = MS_i/MS_E$, where MS_i is the mean square for factor i, with 1 degree of freedom, and MS_E is the error mean square with $n - 1 - k$ degrees of freedom. The LS_i are compared with a critical value from the $F_{[1;n-1-k]}$ distribution, and the dispersion effect of factor i is declared active if LS_i exceeds the critical value; this is a one sided test. This procedure follows the results from Bartlett and Kendall (1946); see also Nair and Pregibon (1988).

Misspecifications in the location model

Pan (1999) shows for the VR method in unreplicated experiments, that $E(\Sigma_j z^2_{j\,|\,i+}) = m_i\sigma^2_{i+} + B(i+)$, and $E(\Sigma_j z^2_{j\,|\,i-}) = m_i\sigma^2_{i-} + B(i-)$, where m_i is defined above, σ^2_{i*} is the mean variance at level * of the ith factor, the bias term $B(i*) = [\mathbf{I}(i*)(\mathbf{Z}\beta_2)]'\,[\mathbf{I}(i*)(\mathbf{Z}\beta_2)]$, $\mathbf{I}(i*)$ is the n x n diagonal matrix with elements 1 in the rows for which factor i is at level *, and 0 elsewhere, $\mathbf{Z}$ an n x p_2 design matrix with non-detected location effects not corresponding to interactions of factor i with detected location effects, and β_2 a p_2-dimensional vector with the corresponding true values. Through the bias terms of the above expected values, the value of F_{VRi} need not be related to the value of $\sigma^2_{i+}/\sigma^2_{i-}$, as indeed is convincingly shown by Pan (1999).

If $B(i+)$ equals $B(i-)$, then the expectation of the VR statistic under the null hypothesis equals 1 (numerator and denominator being independent). Even in this case, the null distribution will not be F; see the subsection on the VR method. However, deviations from the F distribution may not be serious.

To derive practical conditions for equally sized contributions $B(i*)$, consider the dispersion-effect testing of the main effect of factor X. Let there be an undetected main location effect of factor Y. Thus, $B(X+)$ contains accumulated effects of $Y+XY$, while $B(X-)$ contains the accumulated effects of $Y-XY$. It is clear that whenever there are undetected interactions, $B(X+)$ cannot be guaranteed to be equal to $B(X-)$. This is also true if XY corresponds to an undetected main effect, as could be the case in resolution III designs. We conclude that the required conditions are met if the design used has a resolution of IV or higher, while at the same time there are only main location effects.

The LR method in unreplicated experiments will suffer from the same problems as does the VR method. This is because the residual sums of squares modelled prior to construction of the LR statistic equals the sum of the squared contrasts used for the VR method, see Bergman and Hynén (1997). However, due to the logarithmic link function for the dispersion model, in conjunction with additive bias $B(i\pm)$ and the seesaw algorithm, the influence of misspecifications in the location model on the LR statistic is not easily tractable analytically.

A cryopreservation experiment

We will now apply the various algorithms proposed in the previous section to the sub-study of Study 3.2 that assesses the potential of four two-level factors to enhance cell viability after cryopreservation of liver slices. According to the settings of the factor 'thickness' either thick or thin slices of rat livers were cut. The slices were immediately or stepwise immersed in UW or WME as liquid medium according to the settings of the factors 'cryoprotectant' and 'medium', respectively. Subsequently, the slices were either cooled down gradually (slow freezing) or treated with liquid nitrogen (fast freezing)

Table 22. GST activities in the 2^5 sub-study of Study 3.2

experiment	thickness	cryoprotectant	medium	freezing speed	GST activity[a]	
1	thick	immediate	uw	quick	0.0911	0.0772
2	thick	immediate	uw	quick	0.2720	0.2790
1	thin	immediate	uw	quick	0.0760	0.0611
2	thin	immediate	uw	quick	0.4230	0.2950
1	thick	stepwise	uw	quick	0.0482	0.0490
2	thick	stepwise	uw	quick	0.3250	0.2880
1	thin	stepwise	uw	quick	0.0686	0.0582
2	thin	stepwise	uw	quick	0.3150	0.3090
1	thick	immediate	wme	quick	0.1530	0.3390
2	thick	immediate	wme	quick	0.6580	0.4170
1	thin	immediate	wme	quick	0.2798	0.2860
2	thin	immediate	wme	quick	0.5400	0.9730
1	thick	stepwise	wme	quick	0.1961	0.1193
2	thick	stepwise	wme	quick	0.4160	0.4040
1	thin	stepwise	wme	quick	0.1240	0.0904
2	thin	stepwise	wme	quick	0.5160	0.7330
1	thick	immediate	uw	slow	0.0439	0.0495
2	thick	immediate	uw	slow	0.0690	0.1210
1	thin	immediate	uw	slow	0.0644	0.0518
2	thin	immediate	uw	slow	0.1700	0.1180
1	thick	stepwise	uw	slow	0.0510	0.0563
2	thick	stepwise	uw	slow	0.1000	0.1330
1	thin	stepwise	uw	slow	0.0506	0.0511
2	thin	stepwise	uw	slow	0.2870	0.1810
1	thick	immediate	wme	slow	0.1264	0.0707
2	thick	immediate	wme	slow	0.1920	0.1490
1	thin	immediate	wme	slow	0.1881	0.1067
2	thin	immediate	wme	slow	0.1560	0.2050
1	thick	stepwise	wme	slow	0.0683	0.0957
2	thick	stepwise	wme	slow	0.2070	0.1010
1	thin	stepwise	wme	slow	0.0526	0.0726
2	thin	stepwise	wme	slow	0.1730	0.3520

[a]Expressed as μmol of formed product / mg protein / min.

according to the factor 'freezing speed'. The study comprised two experiments separated by a few months. In view of an interest in reproducibility, 'experiment' was considered as a separate factor rather than as a blocking variable. Thus, the relevant design was a 2^5 full factorial.

One of the responses was the activity of the enzyme GST, four hours after thawing of the slices. Table 22 gives the activity for pairs of identically treated slices. The results of the various data-analytic procedures applied to log-transformed GST activities are given in Table 23, grouped according to the number of slices per treatment used for the statistical analysis. The assessment of the location effects based on 2 slices uses a dummy factor 'slice' as well as the interactions of slice with the 5 other factors, to calculate the PSE of 0.059. The 32 replicate variances were judged by the 3 methods to detect dispersion effects listed in the Table. All methods point to a dispersion effect of the factor medium, WME having the larger variance.

The right side of Table 23 gives the results of the testing procedures for unreplicated designs. For this purpose, the 2nd slice of each of the pairs was used. With the PSE of 0.144 the main location effects of experiment, medium, and freezing speed are judged to be active. Using a location model with these factors, both dispersion-testing procedures relevant for unreplicated designs point to the activity of cryopreservation medium.

The PSE for the unreplicated results is more than the expected factor of $\sqrt{2}$ larger than the PSE based on the replicates. This points to the presence of an added variance component between pairs of identically treated slices. Such a component would invalidate the present version of the LR testing, in view of the incompatibility with the logarithmic link function. As a possible solution, we could use an identity link. Note, however, that we use the methods for unreplicated experiments in this context just as an illustration.

Design of the simulation study

We will now simulate the procedures for detecting location and dispersion effects under a wide range of practical conditions. As a vehicle for the simulation, we use the 13 two-level

Table 23. Tests for log-transformed GST activity in the cryopreservation study

factor	2 slices per treatment				1 slice per treatment[b]		
	location effect	F_{LS}	X^2_{LR}	F_{VR}	location effect	X^2_{LR}[c]	F_{VR}[c]
experiment	1.09 (0.059)[a]	0.70	0.34	1.34	1.12 (0.144)	0.03	0.91
thickness	0.18 (0.059)	0.00	0.12	0.84	0.19 (0.144)	0.08	0.85
cryoprotectant	-0.16 (0.059)	1.47	0.30	0.76	-0.14 (0.144)	0.22	1.31
medium	0.61 (0.059)	5.87	5.85	3.48	0.64 (0.144)	7.40	4.83
freezing speed	-0.71 (0.059)	2.83	0.64	1.49	-0.71 (0.144)	1.27	0.54

[a]Figures between brackets: pseudo standard errors. [b]Results for 1 slice based on 2nd slice of each treatment. [c]Dispersion-effect tests for 1 slice per treatment use location model with main effects for experiment, medium, and freezing speed.

Table 24. Experiments considered in the simulation study[a]

N[b]	number of experimental factors					
	4	5	6	7	8	9
16	full	V	IV			
32		full / 2xV	VI / 2xIV	IV		
64			full / 2xVI	VII / 2xIV	-	IV

[a]Replicated designs are indicated with '2x'. [b]Run size.

experiments given in Table 24. The table entries show the resolution of the design(s) considered in Roman numerals. We did not consider resolution III designs in view of the problems with the dispersion-effect detection outlined previously. Replicated designs are marked with '2x'. Thus, for 32 runs and 6 factors we study both the half-fraction of a full design and 2 replicates of a ¼ fraction. The replicated designs were used for dispersion-effect testing only.

Each of the experiments was studied under all combinations of 4 two-level simulation factors. This section summarises the simulation factors, effect sizes, and performance measures of the simulation study.

Simulation factors

Simulation factors, abbreviations of their names, and their settings are given in Table 25. It has been argued in a previous section that the results for dispersion-effects testing are interpretable only in a main-location effects context. This is the low level of the factor 'location effects' (*le*). As the high level of this factor, we included experiments with, for k factors, 1/4 k (k-1) active interactions. This was done, firstly, to extend the study of the performance of location-effects testing, and, secondly, to confirm the expected breakdown of the dispersion-effects testing in the presence of interactions.

From the previous development, we expect to see a difference in performance of the dispersion-effects detection between experiments with just a single and with more than one dispersion effect. For this reason, we included experiments with 1 and those with 3 dispersion effects according to factor n_d.

The nominal Type I error for location-effects detection (α_m) was included in the study to

Table 25. Simulation factors and their settings

simulation factor	abbreviation	low level	high level
location effects	*le*	main effects	main effects + half of interactions
dispersion effects	n_d	1	3
nominal Type I error for location-effects detection	α_m	5 %	1 %
method for dispersion-effects detection[a]	*M*	Likelihood Ratio (LR)	Variance Ratio (VR)

[a]For replicated experiments, ANOVA on logged sample variances (LS) is included as a third level of *M*.

simulate improvement in the location model by including marginally significant effects. This may affect the performance of the dispersion-effects detection.

Finally, the various procedures to detect dispersion effects are also included as a simulation factor (*M*).

Simulated effects

The structure of the simulated data is given in the section on detection of active effects. Experimental factors – as opposed to simulation factors – are indicated with letters A through J (I is omitted). The mean variance of the observations was taken to be 1. Experiments with $n_d = 1$ have a regression coefficient for factor B on the log variance of 0.67. Experiments with $n_d = 3$ had regression coefficients of 0.33, 0.67, and 1, for factors A, B, and C, respectively. These effects correspond with an increase in standard deviation with a factor of 1.4, 2, and 2.7, respectively. Table 26 gives the power of *F* tests for detection of differences by these amounts, for experiments with a single factor. If our detection procedures would resemble ordinary *F* tests, then it is seen that the adopted effect sizes cover a wide range of powers. Even for a 64 run experiment, the smallest effect has less than 50 % chance of being detected, while the largest effect has a fair chance of being detected even with a 16 run experiment.

In all simulated experiments, factors A, B, C, and D have regression coefficients 0.25, 0.5, 0.75, and 1 for the location. Remaining main effects, if any, have regression coefficients of 0.625. If active interactions are simulated, then these have regression coefficients of 0.375. Using the mean variance of 1, the estimators of the location effects are distributed as N (β, $1/\sqrt{N}$). Approximate powers for detection, if σ is assumed to be known, range from 0.16 to > 0.9 for N=16, and from 0.5 to >0.9 for N=64. Thus the smallest effect may just be detected for N=64, but will remain undetected with high probability for N=16. This will realistically confound detection procedures for the dispersion effects.

Performance measures

For each setting of the simulation factors within an experiment, 2500 datasets were generated and analysed with the relevant detection procedures. Thus, fractions of the 2500 datasets in which some effect is detected are correct to 3 decimal places.

For each dataset *i*, detection identifiers β^*_{ij}, $\gamma^*_{ij} \in \{0,1\}$ are recorded, with $i \in \{1 \ldots 2500\}$,

Table 26. Power of *F* tests (%) for three regression coefficients for the logged variance[a]

N[b]	regression coefficient		
	0.33	67	1
16	12	37	69
32	23	71	96
64	46	96	99

[a]Calculations assume a single two-level factor with equal numbers of runs at either level. [b]Run size.

$j \in \{1 \ldots N\text{-}1\}$ (location effects) or $j \in \{1 \ldots k\}$ (dispersion effects). Note that we consider dispersion main effects only.

From the detection identifiers, we construct the following performance measures.

(1) Power for individual location effect β_s: $p_l(s) = \Sigma_i \beta^*_{is}/2500$

(2) Power for individual dispersion effect γ_s: $p_d(s) = \Sigma_i \gamma^*_{is}/2500$

(3) Overall power for location effects:

$$p_l = \frac{\sum_{i=1}^{2500} \sum_{j|\beta_j \neq 0} \beta^*_{ij}}{2500.n\{j \mid \beta_j \neq 0\}}$$

(4) Overall power for dispersion effects:

$$p_d = \frac{\sum_{i=1}^{2500} \sum_{j|\gamma_j \neq 0} \gamma^*_{ij}}{2500.n\{j \mid \gamma_j \neq 0\}}$$

(5) Type I error for location effects:

$$q_l = \frac{\sum_{i=1}^{2500} \sum_{j|\beta_j = 0} \beta^*_{ij}}{2500.n\{j \mid \beta_j = 0\}}$$

(6) Type I error for dispersion effects:

$$q_d = \frac{\sum_{i=1}^{2500} \sum_{j|\gamma_j = 0} \gamma^*_{ij}}{2500.n\{j \mid \gamma_j = 0\}}$$

with $n\{.\}$ denoting the number of elements in the set $\{.\}$. All powers and Type I errors relate to a nominal value of 5%.

For all 9 unreplicated experiments, and all 16 combinations of the simulation factors' settings, we recorded the above performance measures. For replicated designs, we note that the detection of active dispersion effects will neither be affected by the number of active location effects, nor by α_m. We nevertheless completed all 16 combinations involving the LR and VR method. The evaluation of the LS method was based on 2 additional sets of 2500 simulations for 1 and 3 active dispersion effects, respectively.

Simulation results

Null simulations

The detection of dispersion effects is linked to the detection of location effects only for the unreplicated experiments. For these experiments, we simulated the Type I error of the dispersion-effects detection under absence of location as well as of dispersion effects at a nominal value of 5%. Note that, in absence any effect, the Type I error for the location-effects detection is equal to the nominal value. The results of the null simulations are given in Table 27. They are based on 2500 datasets for each table entry. The simulated Type I errors are close to the nominal value. In general, error rates obtained with α_m= 1% are smaller, and closer to nominal value, than the other results. As a possible clue, we note that the residuals considered for the dispersion-effects detection will be smaller for α_m= 5% than for α_m = 1%. This may enhance detection of spurious dispersion effects, because the

relative impact of a large residual will be more pronounced when other residuals are small.

Simulations with active effects

Location effects; Type I error. The logits of the simulated Type I errors were used as responses to evaluate the 15 Yates effects of the 4 simulation factors for each of the experiments. Effects were judged with the help of their PSE. Tables of backtransformed means were constructed, based on the active effects. These tables relate just to the values for α_m=5%. They are given in the left half of Table 28. The results clearly show a conservatism of the PSE tests if the number of active effects approaches, or exceeds, 0.5.(*N*-1). This is because, for these cases, the median of the absolute effects, on which the PSE is based, is likely to be active, resulting in an inflated PSE. We have indicated the cases for which the actual Type I error is less than 1 %.

In the 2^4 and 2^{5-1} experiments, the number of dispersion effects marginally affects the Type I error, but this is not of practical significance.

Location effects; power. The effects of the simulation factors on the power of the location tests were detected in the same way as were the effects on the Type I error. Detection was based on the overall power. Tabulated summaries, however, are based on the effects of experimental factors A, B, C, and D, sized 0.25, 0.5, 0.75, and 1.0, respectively. The right side of Table 28 presents the ranges of simulated powers for α_m=5%. Only the number of location effects appreciably affects the results. We have indicated by boldfacing the cases with power > 70%. These powers are designated acceptable.

For the 16 run experiments, we see that there is little hope for experiments with 6 or more active effects. The 16-run main effects only experiments with 4 or 5 factors have reasonable power only for the largest simulated effect. For the 32 run experiments, effects sized 0.75 and 1 are well detectable, except for the case with 7 factors and 17 active effects, because of the inflation of the PSE. This case has still an acceptable power for effects sized 1. For the 64 run experiments, effects from 0.5 onward have a quite high power, except for the 9 factors/27 active effects case for which, again, the PSE is likely to be inflated. This case has acceptable power for effects larger than 0.75.

Table 27. Simulated Type I errors (%) for two methods to detect dispersion effects, under absence of any active effect[a]

α_m	M	N[b]		
		16	32	64
1%	VR	4.72	5.32	4.96
1%	LR	5.12	5.00	4.72
5%	VR	7.24	5.72	6.84
5%	LR	7.44	5.56	5.00

[a]Type I Errors Have Nominal Values of 5%. [b]Run size.

Dispersion effects; Type I error. The Type I error for the dispersion effects was studied in the same way as the Type I error for the location effects. The results for unreplicated experiments, classified according to the influential effects of the simulation factors, are given in the left part of Table 29. We found no appreciable influence of *M* for any of the experimental configurations studied. That is, the Bergman and Hynen (1997) VR method has Type I errors comparable to those obtained with our LR method.

Table 28. Rejection rates (%) for location-effects detection with nominal Type I error of 5%[a]

design	Type I error			power				
	le			*le*	regression coefficient			
					0.25	0.5	0.75	1
2^4		$n_d = 1$	$n_d = 3$					
	4	2	2	4	7-10	29-31	56-60	**78-82**
	7	0.8	0.7	7	3-4	14-17	39-40	65-66
2^{5-1}		$n_d = 1$	$n_d = 3$					
	5	2	1	5	6-7	24-25	49-51	**73-75**
	10	0.2	0.2	10	0.2-0.4	3-5	13-19	35-38
2^{6-2}	6	0.8		6	3	18-19	38-41	64-66
	13	0.07		13	0	0.6-0.8	5-6	16-17
2^5	5	4		5	21-23	68-70	**94-95**	**99-100**
	10	1		10	10-11	47-50	**85-86**	**98-99**
2^{6-1}	6	3		6	19-30	65-67	**93-94**	**99-100**
	13	0.3		13	4-5	30-32	**70-74**	**95**
2^{7-2}	7	3		7	20	61-63	**91-93**	**99**
	17	0.4		17	0.3-0.8	9-10	39-41	**78-79**
2^6	6	4		6	45-47	**95-96**	**100**	**100**
	13	2		13	37-40	**92-93**	**100**	**100**
2^{7-1}	7	4		7	44-45	**95-96**	**100**	**100**
	17	2		17	30	**87-89**	**100**	**100**
2^{9-3}	9	4		9	44-47	**94-96**	**100**	**100**
	27	0.1		27	6-7	57	**95-96**	**100**

[a])Powers > 70% are in **bold**.

The simulated 2^{9-3} design is the only instance in which α_m affects the Type I error for the dispersion-effects detection. The Type I error for the dispersion effects increases with decreasing α_m. The mechanism may well be that an increase in α_m decreases the number of undetected location effects. The effect of α_m is clearly visible only if there are active location interactions. This may be because biases of numerator and denominator of the *F*

Table 29. Rejection rates (%) for dispersion-effects detection in unreplicated experiments with VR and LR method and nominal Type I error of 5%[a]

design	Type I error			power			
	#le			*#le*	regression coefficient		
					0.33	0.67	1
2^4	4/7	7		4	8-13	7-13	9-19
				7	23-40	3-42	27-49
2^{5-1}	5	3		5	1-6	0.8-10	2-6
	10	9		10	23-27	22-31	19-25
2^{6-2}	6	3		6	4-8	1-8	2-9
	13	*39*		13	52-56	49-65	63-69
2^5		$n_d = 1$	$n_d = 3$				
	5	5	*10*	5	13-17	19-33	47-62
	10	*10*	*11*	10	27-38	37-50	24-34
2^{6-1}	6/13	8		6	4-12	14-25	34-49
				13	20-53	14-31	9-33
2^{7-2}	7	5		7	2-17	6-28	14-49
	17	*15*		17	15-37	15-45	7-52
2^6		$n_d = 1$	$n_d = 3$	6	26-34	**64-80**	**93-97**
	6/13	7	*13*	13	30-41	54-63	**74-88**
2^{7-1}		$n_d = 1$	$n_d = 3$				
	7	8	*13*	7	29-30	**67-75**	**93-96**
	17	6	*10*	17	18-27	41-46	52-64
2^{9-3}		$n_d = 1$	$n_d = 3$				
	9	7	*(11,12)*[b]	9	21-29	59-70	**89-91**
	27	*(19,36)*[b]	*(14,27)*[b]	27	30-71	42-84	41-77

[a]VR and LR methods do not differ; powers with median > 70% are in **bold**. Type I errors ≥ 10% are in *italics*. [b]For α_m of 5% and 1%, respectively.

tests are equal in the main-location-effects only case. The fact that there is a marked influence of α_m with a good number of location effects present extends the results of Pan (1999).

The combined influence of *le* and n_d on the Type I error of the dispersion-effects detection varies over the experiments considered. The activity of these simulation factors is due to the biases in the expected values of the variances and to the correlation of null contrasts. We note that, for a given number of factors, n_d is active only for the larger designs.

We consider tests with a Type I error ≥ 10 % when the nominal value is 5% unfit for practical use. In Table 29, the corresponding sections are italicised. Our general expectation that Type I error is best with main-location-effects only and a single dispersion effect is confirmed.

Dispersion effects; power. The right part of Table 29 summarises the results for the power of the dispersion-effects detection. The results are split up according to the levels of *le*, as this simulation factor induces differences in overall power for all the experiments. The columns in the sub-tables correspond with the small, medium and large dispersion effects. The cells give ranges of 4 (small and large dispersion effect) or 8 (medium dispersion effect) detection probabilities. Effects on overall powers (not shown in the Table) are as follows. For the 16-run experiments, and the 2^6 experiment, *le* is the only discernible simulation factor affecting the overall power. The value for α_m affects overall powers for all 32-run experiments, and the 2^{9-3} experiment. The simulation factor n_d has an influence in the 2^5, and in the 2^{7-1}. Finally, the 2^{7-2} experiment is the single instance for which there is an influence of *M*. LR tests here have more power than VR tests.

We will not try to find explanations for the various effects of the simulation factors. Instead, we will concentrate on the practical feasibility of the testing procedure. The bold-faced parts highlight instances for which there is a power > 70%. It is evident that there is no hope to detect even a 2.7 fold increase in standard deviation for experiments with 16 or 32 runs with the present procedures. For the 64 run experiments, we can detect a doubling of the standard deviation in the main-location-effects-only experiments with 6 or 7 factors, and a 2.7 fold increase in the experiment with 9 active main effects. The 6-factor experiment also permits detection of the large dispersion effect if there are active interactions. However, we do not recommend the tests for this case, as we believe that the results depend too much on the particular interactions that are active.

Note that the 64 run experiments must have at most one dispersion effect. Otherwise, the Type I error is too high. This seemingly compromises the results on the large dispersion effect, as these are obtained only for cases with $n_d = 3$. We believe, however, that the bold-faced power reported for the large dispersion effect also applies if this were the only dispersion effect present. This is because the power for the medium dispersion effect in the 1-effect case is not less than the power for the 3-effects case.

Dispersion effects; replicated designs. The obvious advantage of using replicated designs for the detection of dispersion effects is their insensitivity to the number of active location effects. Table 30 gives the rejection rates for the simulations. As the active dispersion effects are the same in each of the experiments, the results for equally sized experiments are virtually identical. For this reason, we report the results according to the total number of

runs. We see that the LS method has a good control of the Type I error, even for cases with 3 dispersion effects. The VR and LR methods have an acceptable Type I error only if there is just a single active dispersion effect.

The power of LS tests is much lower than that for the VR and LR methods. Unlike their unreplicated equivalents, the replicated 32-run experiments permit detection of a single large dispersion effect with the LR or the VR method, but not with the LS method. For the replicated experiments with 64 runs, the VR and LR methods can detect moderate or large dispersion effects, and the LS method can detect large dispersion effects. The power of the VR and LR tests here is comparable to the power for the 64-run unreplicated experiments with main-location-effects only.

Discussion

This study uses standard two-level fractional or full factorial designs to detect both location and dispersion effects. For the detection of location effects in unreplicated experiments, we conclude that the PSE remains of practical use even in the presence of several dispersion effects. This extends the finding of Pan (1999) for location-effects detection in the presence of a single dispersion effect.

Using a quantile q <0.5 as a basis to calculate a PSE (Haaland and O'Connell, 1995) may increase robustness against active location effects. A referee of this study suggested the use of more efficient location-effects detection methods than the PSE. We note that such a method would typically involve a maximal location model. The section on location-effects detection gives an argument against an approach with such a model for screening designs. However, if the experimenter is willing to adopt resolution IV where resolution III would suffice to detect location effects, one may well adopt a maximal location model. Thus we

Table 30. Rejection rates (%) for dispersion-effects detection in replicated experiments with nominal Type I error of 5%[a]

N[b]	Type I error			power			
	M			*M*	regression coefficient		
					0.33	0.67	1
32		$n_d = 1$	$n_d = 3$				
	VR	8	*16*	VR	22-24	41-46	**71-73**
	LR	9	*17*	LR	23-25	43-47	**73-76**
	LS	5	4	LS	9	22	42
64		$n_d = 1$	$n_d = 3$				
	VR/LR	9	*18*	VR/LR	30-33	**68-77**	**94-95**
	LS	5	5	LS	15	40	**70**

[a]Powers with median > 70% are in **bold**. Type I errors ≥ 10% are in *italics*. [b]Run size.

feel that this topic is an attractive possibility for further research.

This study did not include location-effect detection for replicated experiments, as our focus was primarily on the use of screening designs in detecting dispersion effects. Replicated designs were used to study the more direct assessment of dispersion effects by using replicate variances. Lee and Nelder (1998) suggest to detect location effects in replicated experiments by weighted regression, using the inverse of the modelled variances as weights.

We considered dispersion effects that are much smaller than those featuring in Bergman and Hynén (1997) and Pan (1999). We did so, because we were originally intent of studying for our clients from industry possibilities of the usual two-level designs to detect a doubling of the standard deviation. Of course, larger dispersion effects will be detected more easily.

Our simulations for dispersion-effects detection in unreplicated experiments suggest that the VR and LR methods perform equivalently in dispersion-effect detection. We think that the results for the VR method are conclusive. However, for the LR method implemented in this study, results will also depend on the strategy used to employ the method. We chose to let the LR tests mimic the VR tests through testing the experimental factors one at a time with reference to a null dispersion model. We could consider these test results as necessary for a first step in a forward selection procedure, but we did not record the effect to be eventually selected (we will remedy this in a future version of this study). Possibly, effects not detected in the first step could be detected in subsequent steps. Powers for the effects that should be selected in the first step in 16 or 32 run experiments are quite low, however. Thus, there dose not seem to be much hope that the active effects may be detected in further steps of a forward selection procedure.

A more comprehensive study would also consider the results of a whole sequence of forward selection tests, or even backward elimination in a main-effects model. As these additional issues would permit a more definitive assessment of LR dispersion tests, this is a very important subject of further research. We note, however, that this kind of studies typically involve a heavy amount of computing time. This may well be the reason why, to our knowledge, this kind of studies has not yet appeared in the literature.

Dispersion-effects detection in unreplicated experiments with the VR method heavily relies on a main-location-effects only assumption. This result is based on theory; see the subsection on misspecifications in the location model. It is plausible that the theoretical result will also hold for our LR method. Practical results on violation of the main-effects only assumption were obtained with the computer simulations. However, a referee pointed out that simulations with interactions are also those with a large number of location effects. This is true for the 16-run experiments. We note that for this run-size Pan (1999) already provides extensive support for our statement. For 32-run and 64-run experiments we note that our full factorial designs with active interactions could be considered as experiments with a moderate number of active effects. For the 32 run full factorial, 10 out of 31 effects were active, while for the 64 run full factorial, 13 out of 63 effects were active. However, one could wish that there were experiments with some active main effects replaced by interactions of the same size. For a future version of this study, we plan to do simulations with 32 runs and 5, 6, or 7 factors, with some main effects replaced with interactions. We plan to do the same for 64 run experiments and 6, 7 or 9 factors. In absence of the ensuing results we may point out that the 64 run experiments with 17 active effects shows a

dramatic drop in power for the dispersion effects when compared to the experiment with 13 active effects, while the 17-effects experiment has just a single additional main effect and three additional interactions.

For the VR method in unreplicated experiments, the following conditions have to be jointly met for a safe detection of a dispersion effect doubling the standard deviation: there is at most a single dispersion effect; the run-size is 64 or larger; the resolution is larger than III. The performance might be improved by including all main effects in the location model instead of just the detected ones. This removes the bias from the expectation of the sums of squares. However, the distribution of the VR will still not be F, and the main-location-effects assumption cannot be discarded.

The LR method in unreplicated experiments, applied with n_d=3 as well as with n_d=1, does not permit a sensitive detection of the largest effect (this is the only effect for n_d = 1) in the first step of forward selection in 16-run and 32-run experiments. Experiments with 64 run experiments have sufficient power for medium and large dispersion effects. However, more work is needed for a definitive assessment of this method.

For replicated experiments, we compared the VR method with an LR method using a GLM on sample variances and with the LS method. The LS method simultaneously fits all dispersion main effects. As the simulated designs are orthogonal, the LS results permit assessment of a final dispersion model. We also tried to use a maximal dispersion model for the LR method. This was a failure, however, as the algorithm failed to converge in a substantial fraction of the simulations. This could be due to the fact that the replicate variances have just a single degree of freedom.

Our simulations show both VR and LR methods to be slightly progressive when there is a single dispersion effect, and heavily so when n_d = 3. The LS method has a good control of the Type I error at the cost of having far less power than the other methods considered here. We conclude that, for replicated designs, we should use the VR method only if at most a single dispersion effect is suspected, and the LS method when there could be more than 1 dispersion effect active. More work, notably on effects actually selected after the first step in a forward selection, is needed to interpret the results for the LR method.

The less problematic character of the dispersion-effects detection in replicated experiments is counterbalanced by the loss in resolution for location effects. If we wish to gain in resolution by using unreplicated experiments, then we have to adopt a main-location-effects-only assumption to be able to detect dispersion effects safely, at least with the VR method and the present version of the LR method. However, we could then use a replicated resolution III design! Therefore, we don't recommend the VR method or our LR method in unreplicated two-level experiments with the purpose of detecting dispersion effects of the size considered in our study.

Chapter 4 Statistical designs in toxicology of mixtures

4.1 Statistical designs in combination toxicology: a matter of choice

Introduction

The emphasis on statistically designed experiments at the European Conference on Combination Toxicology in 1996 demonstrates an increasing awareness of their efficiency. Indeed, designed experiments meet the needs of a combination-toxicologist to study the effect of several compounds in one experiment. Of particular use here are factorial designs (see General Introduction) and designs for response surface analysis (Box and Draper, 1987, pp. 502-524).

There are three reasons why the above designs enhance efficiency of research. Firstly, they enable economy of experimentation through the use of fractional factorial designs (Box *et al*., 1978, pp. 374-418), in which only a fraction of a full factorial design is run. The usefulness of such a design was shown experimentally by Groten *et al*. (1991), who investigated the influence of seven minerals with eight experimental groups.

A second efficiency enhancing feature of designed experiments is their robustness against inhomogeneity of experimental material through the use of blocking (Cox, 1958). This device works well, because the comparisons of interests are carried out within homogeneous groups of experimental units called blocks. Block differences, therefore, do not disturb the comparisons. A clear experimental demonstration of the use of blocking is given by Littlefield *et al.* (1980a, 1980b), who reported on the induction of liver and bladder tumours resulting from exposure to 2-acetylaminofluorine. The animals of the top shelf of cages had increased food consumption and a lower weight-gain than animals on lower shelves. This did not affect, however the comparisons of interest because the shelf was a blocking factor: each shelf had cages of all treatments.

Finally, factorial experiments yield parameter information of high precision because they use the whole of the data for calculating the effects. Groten *et al*. (1997; Study 4.2) demonstrate the use of a two-level experiment on additivity of nine compounds in sixteen experimental groups of five animals. Effects are calculated as the difference of two means of 40 animals each.

Multifactor experimental plans designed specifically for exploring response surfaces do not have a factorial structure. They nevertheless yield high precision parameter information, because the experimental points have a well-determined spread. Box and Hunter (1957) theoretically illustrate this notion. Svendsgaard and Hertzberg (1994) and Carter and Gennings (1994) mention the above experimental plans as an option for toxicological experiments; Study 4.3 presents a real experimental illustration of such designs.

One of the lesser-known aspects of experimental designs is that there are often several ones to meet the purpose of the study. It then becomes a matter of debate which of the designs to choose.

It is the purpose of this study to demonstrate the process of choosing between alternative designs in conjunction with the above efficiency-enhancing issues. Three experiments are used for illustration. The first studies the combined effects of nine compounds. It consists of 16 experimental groups of five rats. Two levels are used for each compound. Using 32 experimental groups of two animals gives a more straightforward interpretation of the effects, at the cost of higher complexity in experimental management. The second experiment deals with *in vitro* testing of mixtures of three aldehydes, each studied at three levels, with an additional control group. The problem here is to divide 28 treatments over nine plates of 12 wells each. There is a trade-off between the degree of protection against irreproducible effects of plates and complexity of the experimental design. Finally, a sensory irritation study with mixtures of three aldehydes in rats is used to illustrate choosing between several designs for response surface analysis.

This study has a separate section for each of the above experiments. The final section contains some concluding remarks.

Economy of experimentation

Study 4.2 demonstrates the detection of interactive effects using a two-level factorial design. The chemicals under investigation are Sn, Cd, butyl-hydroxyanisol, di(2-ethylhexyl)phtalate, methylene chloride, spermine, loperamide, aspirin, and formaldehyde. Each compound was either administered at the minimal adverse effect level (MAEL) or not at all.

A full two-level design will take $2^9 = 512$ treatment groups. With such a design the following 511 effects may be calculated for each study parameter. In the first place there are nine main effects. These are differences between the mean of 256 groups with a certain compound at its MAEL, and the mean of the remaining groups in which the compound is absent.

A second group of effects consist of 36 first-order interactions. These are calculated as the difference between a main effect of compound 1 when compound 2 is also present and the main effect of compound 1 in the absence of compound 2. This difference is then divided by 2. Therefore, first- order interactions, like main effects, are differences between two means of 256 groups each.

Finally, by analogy with first-order interactions, second-order up to eighth-order interactions, namely interactions between two to nine factors, can be estimated. Note that the term ‘interaction’ in a two-level context describes non-additivity of effects in the sense of lack of effect-addition. A reference to full two-level designs is Box *et al*. (1978, pp. 306-351).

Investigation of combined effects of nine compounds with 256 experimental groups is as impractical as it is unnecessary. In particular, if one is prepared to assume, at least tentatively, that the degree of non-effect-additivity between two compounds is not altered by any other compound, then we do not need the interactions of order two and above.

Statistical less-than-full designs are called fractional factorials (see Box *et al*., 1978, pp. 374-418). Two of these which are considered here are the 1/32 fraction with 16 experimental groups of four animals each and the 1/16 fraction with 32 experimental groups of two animals each. Each design has a total of 64 animals. It is obvious that both offer quite a reduction in the number of treatment groups. The price to be paid for this, however, is the aliasing (confounding) of effects. Table 31 compares the aliasing in each of the fractions, under the described tentative assumption.

From Table 31, we see that each 'effect' in the 1/32 fraction is either the sum of a main effect and one or four first-order interaction, or the sum of four first-order interactions. The entangling is less severe for the 1/16 fraction. All main effects are free from first-order interactions. Also, the first-order interactions with E are free of other first-order interactions and of main effects. Finally, there are 12 pairs and one quadruplet of first-order interactions; there is one quadruplet of supposedly negligible second-order interactions free from main effects or first-order interactions.

At first sight, the entangling for the 1/32 fraction may seem too severe, because main effects are aliased with first-order interactions. The severity may be relieved, however, by choosing for factor J formaldehyde. This is a compound with very specific target-organs. Ambiguity between main effects of the remaining compounds and interactions with formaldehyde may be solved with this measure.

As far as clarity of the effects is concerned, the 1/16 fraction is superior to the 1/32 fraction. There were, however, two reasons to favour the more fractionated design. First, the experimental conduct for a study in which there are gaseous compounds as well as solid or fluid ones is very complex. 16 different treatment groups already make a heavy demand

Table 31. Main effects and first-order interactions of two fractions of a 2^9 design[a]

1/32 fraction	1/16 fraction
A + FJ	A,B,C,D,E,F,G,H,J
B + GJ	AC + FH
C + HJ	AD + FJ
D + EJ	AF + BG + CH + DJ
E + DJ	AB + FG
F + AJ	AG + BF
G + BJ	AH + CF
H + CJ	BJ + DG
J + DE + AF + BG + CH	AJ + DF
AB + CE + FG + DH	CD + HJ
AC + BE + DG + FH	BC + GH
AD + EF + CG + BH	BD + GJ
AE + BC + DF + GH	CJ + DH
AG + CD + BF + EH	BH + CG
AH + BD + CF + EG	AE,BE,CE,DE,EF,EG,EH,EJ

[a]Single letters denote main effects of factors; two-letter words are first-order interactions. Effects separated by comma's are not aliased with any main effect or first-order interaction; the 1/16 fraction has an additional effect corresponding to AEF + BEG + CEH + DEJ.

on experimental management. 32 groups are forbidding from this point of view. Secondly, with an equal total number of animals, the sixteen-groups design gives more information on the uncontrolled variation. This is particularly interesting for detecting inhomogeneity in the response to a compound, see Study 3.3. The reasons mentioned decided the choice in favour of the 1/32 fraction. The eventual number of animals, five per group, was chosen for convenience rather than out of necessity: this is the usual number per case.

Robustness against inhomogeneity of material

Cassee *et al*. (1996a) studied the cell-damaging capacity of formaldehyde, acrolein, and crotonaldehyde in an *in vitro* experiment. The three aldehydes were studied at three levels. All 27 combinations were to be tried.

The study was carried out by incubating cells with one of the combinations or a control. Cell-damage was measured by the amount of extinction relative to the control. The cells were incubated in plates with twelve wells, arranged in three rows by four columns. There were nine plates at our disposal. The problem is to divide the treatments over the plates.

The above problem is clearly not of the same type as in the previous section. Here the problem is one of blocking the treatments. In solving the problem, the author first looked at the control and then coped with the combinations of aldehydes.

One reason for incorporating controls in any study is to ensure comparability between the studies. On a smaller scale, controls are used to eliminate possible differences between plates. This means, of course, that each plate has to have at least one control. The number of three controls was chosen because this leaves nine wells per plate for the 3^3 part of the study.

An implicit assumption when using controls on each plate is that the ratio of 'extinction for factorial treatment *t*' to 'extinction for control' does not depend on the plates used. Equivalently, the difference in logarithms between the extinction for treatment *t* and for the control should not depend on the plates. The difference in logarithms is statistically easier to deal with, because the usual analysis of variance models assume that extinctions can be predicted by addition, as opposed to multiplication, of various contributing sources of variation.

Ideally, an experiment should be such that assumptions such as the above one can be checked, and that there is an alternative strategy of analysis when the assumptions are violated. In the present case, for example, the difference in logarithms could be inconstant because a factorial treatment may be affected quite differently by small variations in incubation circumstances than a control. If such behavior is detected, then an alternative strategy of analysis could be that only the information of the treatments themselves be used. This would mean that the influence of increase or decrease of the three aldehydes could still be quantified without reference to a zero-damage level.

Division of the 27 factorial treatments over the available nine plates with nine wells was now undertaken. A standard way of doing this is to use three replicates of a 3^3 design confounded in three blocks of nine units each (see Montgomery, 1991, pp. 402-403). Within each plate the treatments should be randomised over the wells.

An alternative that is much more feasible practically is to treat each of the three wells of a column alike. We then have to resort to a single replicate of a 3^3 design laid out in nine

plates of three columns each (see for a general method Montgomery, 1991, pp. 404-405). Here, the treatments should be randomised over the columns.

After completion of the experiment, logarithms of each extinction may be taken. Subsequently, the means of the three controls are calculated for each plate. These are subtracted from the log extinctions of the remaining wells on the corresponding plate. The analysis of variance for the quantities thus calculated is outlined, for both alternatives, in Table 32.

Both designs in Table 32 have random and systematic effects between the plates. The first alternative also has both random and systematic effects for which plate differences cancel out. These effects have a random component due to uncontrolled variation between wells. The second design has random and systematic effects in which column means are involved. Finally, there are effects in which the column means cancel out. These are subject only to random variation between the wells.

A feature of the triplicated three blocks of nine units design is that part of the second-order, or three-factor, interaction between the aldehydes is confounded with plate differences (see Table 32). The remaining plate differences are due to random differences in material or incubation conditions, while there is also a random difference caused by the various wells of a plate. The random effects caused by plates plus wells can be tested by comparing with the random error between wells within the plates.

If plate differences with respect to differences in logarithm are detected, then violation of the usual assumption regarding controls is observed. We could then proceed with just the

Table 32. Analysis of variance for two distributions of 3^3 treatments over 9 plates of 9 wells[a]

alternative	source of variation	df[b]
three replicates	**between plates**	
of 3 plates	second-order interaction + *plates* + *wells*	2
of 9 wells	*plates* + *wells*	6
	between wells	
	main effects + *wells*	6
	first-order interactions + *wells*	12
	second-order interaction + *wells*	6
	wells	48
one replicate	**between plates**	
of 9 plates	first-order interactions + *plates* + *columns* + *wells*	6
of 3 columns	second-order interaction + *plates* + *columns* + *wells*	2
of 3 wells		
	between columns	
	main effects + *columns* + *wells*	6
	first-order interactions + *columns* + *wells*	6
	second-order interaction + *columns* + *wells*	6
	between wells	
	wells	54

[a]Random effects are printed in italics. [b]Degrees of freedom

logarithm of extinctions for the factorial treatments. Even here, the assumption of equality of treatment effects over plates could be roughly checked, because there are three plates with identical treatments.

It is much more difficult to check the assumptions for the design in which the wells within a column are treated identically. It can be seen from Table 32 that random differences between plates can be checked if it is assumed that there is no second-order interaction. The check can then be carried out by comparing the random variation between plates with the random variation between columns within plates. However, absence of a second-order interaction can only be checked if there are no random differences between columns.

While it is not unusual to assume that there is no second-order interaction, a further problem of the nine plates of three columns design is that the random error between columns is estimated with much less degrees of freedom than the random error between wells in the alternative design. This design is therefore less sensitive in picking up treatment effects if the columns do have a random contribution. It is, however, capable of checking plate differences and column differences under absence of second-order interactions.

The aforementioned statistical considerations bear to a heavy extent on checking assumptions on material and conditions that are readily adopted by many toxicologists. Eventually, therefore, the nine-plates three-columns design was chosen for experimental plan. This design offers a combination of simplicity in experimental conduct and capability of checking, at least partially, the properties of experimental material and conditions.

Precision of parameter information from response surface designs

A common device for the study of drug interaction is the isobolographic method (Gessner, 1988). This method attempts to find sets of mixtures of drugs eliciting the same degree of response from experimental animals. For two drugs, the sets may be visualised as a line in an XY plane, the axes denoting the concentrations of the drugs. Such a line is called an

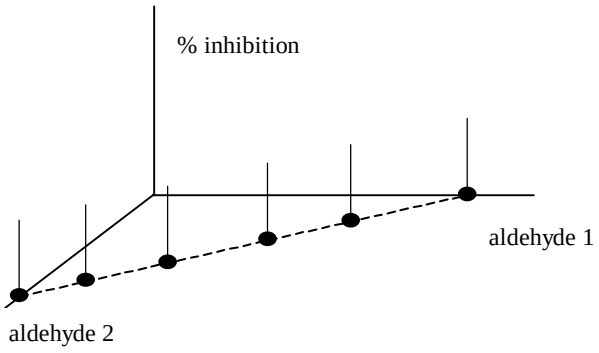


Fig. 7. Spread of experimental points for determination of an isobole

isobole. The point of interest lies in whether the isobole is a straight one, connecting the concentrations of unblended drugs with the same degree of response, or a curved one. The former case is a specific form of dose-addition: the concentration of one of the drugs can be considered here as a dilution of the other one. The latter case may reflect other kinds of addition, or interaction (Steel and Peckham, 1979).

An alternative strategy to determine isoboles is to make up an equation expressing the dependence of the degree of response to the drugs administered. Such an equation, for compounds F, and C, respectively, may take the form % response = $c_0 + c_1F + c_2C + c_{11}F^2 + c_{22}C^2 + c_{12}FC$, or indeed any part of this.

Isoboles can be calculated by equating the right-hand side of the formula to a fixed value of the response and expressing one component as a function of the other. It is immediately clear that this is more universal than determining a single isobole because, first, the approach can easily be generalized to more than two drugs, and, secondly, more than one isobole can be determined with a single parental equation of the aforementioned type.

The experimental points needed to determine the above equation in the XY plane substantially differ from those needed for drawing up a single isobole. The latter points commonly lie around the hypothetical line of additivity. A somewhat exaggerated impression of this is given in Fig. 7. If these points are used to draw up the equation given above, then the coefficients are determined very imprecisely. This comes as no surprise, because the experimental points inadequately support the response surface above the XY plane that is represented by the equation. There is therefore a wide range of values for the slope of the surface compatible with the experimental points.

An experimental arrangement more capable to support a response surface is shown in Fig. 8. The points will yield more precise coefficients, because they are much more spread-out than those of Fig. 7. How exactly the points should be located has been the subject of much research in the area of response surface analysis. A general reference is Box and Draper (1987).

In an *in vivo* study carried out by Cassee *et al.* (1996a), the percentage inhibition of respiratory frequency through exposure to mixtures of formaldehyde, crotonaldehyde and

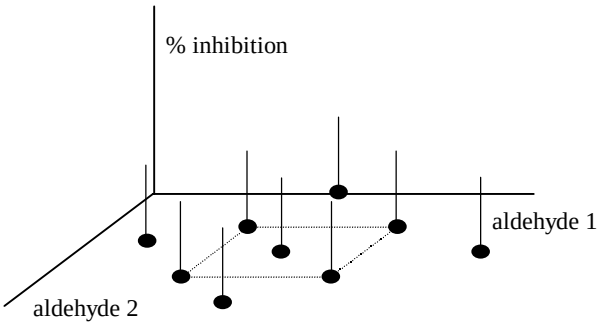


Fig. 8. Spread of experimental points for determination of a response surface

acrolein was to be studied with a design for response surface analysis. Now the 3 x 3 x 3 factorial design could be used for the purpose. There are, however, multifactor experimental plans designed specifically for the exploration of response surfaces involving a lower number of experimental points. The two types used most frequently are the central composite design (Box and Wilson, 1951) and the Box-Behnken design (Box and Behnken, 1960). Though originally made for the empirical search for optimum conditions, these designs can also be used to explore a limited region of the experimental area (Box and Draper, 1987, p. 17).

Fig. 9 shows central composite and Box-Behnken designs for three experimental variables. Both designs aim to reproduce a response surface governed by

% inhibition = $c_0 + c_1F + c_2C + c_3A + c_{11}F^2 + c_{22}C^2 + c_{33}A^2 + c_{12}FC + c_{13}FA + c_{23}CA$.

In this equation, C, F, and A denote the concentrations of crotonaldehyde, formaldehyde and acrolein, respectively. The coefficients are denoted by subscribed c's.

The Box-Behnken design uses 15 points and three different concentrations per variable (there are three points in the centre of the design). A comparable central composite design has 17 experimental points (three in the center) and a total of five different concentrations per variable. Therefore the latter design is more laborious in use than the former.

Both designs allow for the checking of the model, because the experimental points need to estimate only nine coefficients. For an example of model checking for the central composite design, see Box and Draper (1987, pp. 457-461). The appropriate reference for Box-Behnken designs is Draper *et al*. (1994).

A feature of the central composite design that is not shared by the three-variable Box-Behnken design is its capability for blocking (Box and Hunter, 1957). The experimental groups can be divided in two sets of five groups and one set of seven without affecting its ability to estimate the influence of the aldehydes. This can be a distinct advantage when we have a limited number of animals available per period of investigation.

The design eventually chosen was the Box-Behnken one. Apart from a smaller total number of animals, the smaller number of distinct concentrations to be used decided the issue. This number is important, because it was hard to attain prespecified concentrations. In the course of the experiment, it became clear that the control over the concentrations was

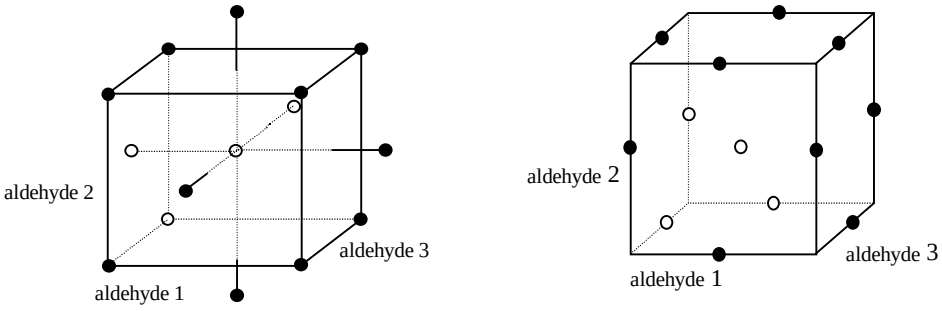


Fig. 9. Experimental points for central composite design (left) and Box-Behnken design (right) for three experimental variables.

insufficient to attain the experimental points as required by the design. The set of experimental points finally reached thus does not exhibit the Box-Behnken structure. Therefore, the safeguards of the design originally intended could not do their job. In particular, curvature-describing terms due to one component now are partly entangled with interactions - in the sense of no effect addition - between two components, while at the same time the effects of the components are estimated with varying degrees of precision.

Fortunately, the strive to attain the experimental concentrations for the Box-Behnken design resulted, in this case, in a set of experimental points which were spread out broadly over the experimental space. General regression techniques (Draper and Smith, 1981), though less clear-cut than those for the Box-Behnken design (Draper *et al.*, 1994), may well be used to model the results. In fact, these techniques established linear and quadratic effects, but no interactions. In retrospect, therefore, absence of the above safeguards did not prevent obtaining valuable results from being obtained.

Conclusion

This article features three efficiency enhancing aspects of experimental design, namely economy of experimentation, robustness against inhomogeneity of experimental material, and increasing the precision of results. To demonstrate these features, the designs of three recent toxicological experiments were evaluated. In the first, effect addition in rats was studied with a fractional two-level factorial design. This type of design affords the study of p compounds, each studied at two levels, in less than 2^p groups. In this case, nine compounds were studied with 16 experimental groups of five animals. There is an alternative design in which 32 experimental groups are used, each group having two or three animals. The author has shown that the statistical basis for evaluating both designs is the aliasing pattern of effects and the precision of the estimates of spread within each group.

In the second experiment, cell damage caused by three aldehydes was studied with an *in vitro* experiment. Three levels for each of the aldehydes were used. The problem here is the division of 3^3 treatments plus controls over nine plates with 12 wells. Two alternatives were shown which differ statistically, first, in their extent of aliasing differences between plates with effects due to the aldehydes and, secondly, in the sensitivity of detecting treatment effects in the presence of inadvertent differences between columns of wells.

The third experiment, on sensory irritation, shows how experimental results obtained with response surface designs may be used to calculate isoboles. The two alternative designs here differ in their capability to make groups of animals handled simultaneously, and in the number of distinct concentrations of the compounds.

The described alternatives differ practically in complexity of experimental conduct. It is concluded that the choice between the alternatives is a result of balancing practical complexity of the design against informational gain. Continuous co-operation between toxicologists and statisticians is needed to strike the balance again and again.

4.2 Subacute toxicity of a mixture of nine chemicals in rats: detecting interactive effects with a fractionated two-level factorial design

Introduction

The bulk of studies to assess the toxicity of chemicals deals with exposures to single compounds (Yang, 1994a). Although toxicity studies of single compounds are important for obtaining basic toxicological information, man is always simultaneously exposed to a large number of chemicals. With the possible exception of some specific mixtures, it is uncertain how the combined toxicity of these chemicals should be assessed or how combined toxicity should be taken into account in standard setting for the individual compounds.

The main problem in the risk assessment of chemical mixtures, are in fact the possible chemical interactions that hamper the prediction of the toxicity of the mixture. Since these interactions may occur at various endpoints in the toxicodynamic as well as the toxicokinetic phase, the combination toxicologist has the unpleasant task to deal with a vast scope of chemical interactions. Because of this multitude of interactions, each possible chemical interaction cannot be tested individually, and due to the immense number of combinations involved, systematic and complete toxicity testing of chemical mixtures is practically impossible. Studies dealing with the toxicology of mixtures are confronted with this problem; for instance Yang and Rauckman (1987) indicated that a 25-chemical mixture of ground water contaminants has 2^{25}-1, or 33,554,431 possible combinations to be tested. A design to test the systemic toxicity of this huge number of groups is virtually impossible from an ethical, economical and practical point of view. One way to overcome this problem is to treat the mixture as a single compound and to test the mixtures as a whole. This approach has been advised for mixtures that are not well characterised (Mumtaz *et al*., 1993), but it has also been applied for assessing the combined toxicity of defined chemical mixtures consisting of nephrotoxicants, pesticides, carcinogens and/or fertilizers (Jonker *et al*., 1993; Charturved *et al*., 1993; Heindel *et al*., 1994; Feron *et al*., 1995a; Ito *et al*., 1995). In these studies, an experimental design was chosen reflecting the net combined effects of all components in the mixture; to limit the number of test groups possible, interactive effects of the components in relation to the effects of individual chemicals were not taken into account.

Whether mixtures should be tested in a simple, defined mode or in a complex, undefined mode, the main question from a governmental perspective is whether the exposure to mixtures of chemicals at low, realistic doses is of real health concern. Although the number of papers dealing with the toxicology of mixtures gradually increases, most of these studies are restricted to two or three compounds tested in short-term studies at relatively high, often toxic doses. There is a clear lack of information on prolonged, repeated toxicity studies on combinations of chemicals (more than two) at low (non-toxic) and sub-toxic doses (Krishnan *et al*., 1991, Heindel *et al*., 1994). In acute and sub-acute toxicity studies in rats, it has been shown that combined oral administration of compounds at the 'no-observed- adverse-effect-level' (NOAEL) of each of them did not lead to clear additivity or synergism of effects, provided the mechanism of action of the compounds was dissimilar (Jonker *et al*., 1990; Jonker *et al*, 1993). In contrast, in a 4-week toxicity study with mixtures consisting of four

nephrotoxicants with similar mode of action, it was shown that the dose-additivity rule could be applied (Feron *et al*., 1995c) at dose levels around the NOAEL. These 4-week toxicity studies were interpreted based on the common approaches in the assessment of mixtures; for a combination of compounds with a similar mode of action one might expect dosis-addition, whereas compounds with a dissimilar mode of action may show effect addition (Bliss *et al*., 1939, Placket and Hewlett, 1952, Mumtaz *et al*., 1994). The results of Jonker *et al*. (1990, 1993) indicate that general application of the additivity rule in risk assessment of exposure to chemicals cannot be justified. Since this preliminary conclusion may have far-reaching consequences for health risk assessment, it was deemed desirable to carry out a similar type of study using a combination of compounds highly relevant to the general population in terms of use pattern and level and frequency of exposure. Such a study should not only be focused on the effect of the mixture as a whole, but also on possible interactive effects between the compounds of the mixture.

One way to detect interactive effects, i.e. non-additive effects, between more than two chemicals in a chemical mixture is to use factorial designs. The use of factorial designs, in which *n* chemicals are studied at *x* dose levels (x^n treatment groups) has been suggested by the US Environmental Protection Agency as one of the valuable statistical approaches for risk assessment of chemical mixtures (Svensgaard and Hertzberg, 1994). In the literature, a 2^5 study was presented to describe interactions between the carcinogenic activity of 5 polycyclic aromatic hydrocarbons (Nesnow *et al*., 1994) and a 5^3 study has been used to identify non-additive effects of three compounds on developmental toxicity (Narotsky *et al*., 1995). Full factorial designs however, lead to very costly experiments, and even if only two dose levels are used, it is already virtually impossible to perform complete, conventional toxicity tests using 2^n test groups to identify interactions between all chemicals of interest. One way to deal with this problem is the use of fractionated factorial designs. These fractionated designs still identify most of the interactions between the compounds and determine which compounds are important in causing effects, but have the advantage that the number of test groups is manageable (Box *et al*., 1978). Fractional factorial designs have been shown to be an efficient, i.e. cost-effective, approach to identify interactive effects between seven trace elements and the cadmium accumulation in the body (Groten *et al*., 1991) and to determine structure-activity relationships for ten halogenated aliphatic hydrocarbons (Eriksson *et al*., 1991).

The present study was intended to find out whether simultaneous administration of nine compounds at a concentration equal to the NOAEL for each of them would result in an NOAEL for the combination. The second aim was to test the usefulness of fractionated factorial models to predict possible interactions in chemical mixtures without the obligation to test large numbers of combinations.

Materials and methods

Materials

Dichloromethane (methylene chloride), cadmium chloride, di-(2-ethylhexyl)phtalate (DEHP) and acetyl salicylic acid (Aspirin) were obtained from E. Merck (Darmstadt, Germany. Paraformaldehyde was purchased from Janssen Chimica (Beerse, Belgium) and loperamide was generously provided by Janssen

Chimica (Beerse, Belgium). Butyl hydroxyanisol (BHA), Stannous chloride and Spermine hydrochloride were obtained from Fluka (Basel, Switserland). All test materials were of analytical grade.

Animals and maintenance

Albino male rats, Wistar outbred [Crl(WI)WU(BR], were obtained from a colony maintained under SPF-conditions at Charles River Wiga GmbH, Sulzfeld, Germany. At the start of treatment, the rats were 9-10 weeks old. Before the start of the experiment, the rats were housed under conventional conditions in an animal room four or five males to a cage in suspended stainless steel cages fitted with wire mesh floor and front. During the experiment the rats were housed individually in inhalation chambers in similar stainless steel cages. Animals were rotated weekly, in such a way that they were kept equally divided over each chamber. The room temperature was kept at 22 ± 2 °C, and the relative humidity at 40-70%. During the exposure the animals had no access to food or water. After exposure, the animals were fed the Institute's cereal-based rodent diet or one of the test diets. Diets and drinking water were provided *ad libitum.* Community tap water was supplied from an automatic drinking-water system. Lighting was artificial, with a sequence of 12 h light and 12 h dark.

Experimental design and treatments

A 4-week oral/inhalatory study was performed in which the toxicity of combinations of nine compounds was examined. The nine chemicals were selected from the database on 4-week toxicity studies previously performed at our institute. The studies chosen were public accessible. The chemicals that were selected represent a combination of compounds highly relevant to the general population in terms of use pattern and level and frequency of exposure (drugs, food additives, food contaminants, biogenic amines, and industrial solvents). The compounds were not chosen based on their similar mode of action. The study comprised 20 groups, four main groups (eight animals/group) and 16 satellite groups (five animals/group).

Main groups. Rats of the main groups were simultaneously exposed to nine chemicals (dichloromethane, formaldehyde, aspirin, DEHP, cadmium chloride, stannous chloride, BHA, loperamide and spermine) at concentrations equal to the "minimum-observed-adverse-effect-level" (MOAEL), NOAEL or 1/3NOAEL; the fourth group was a control group receiving basal diet not supplemented with test chemicals and breathing fresh air. The dose levels were established based on results of previous 4 wk-studies, which were performed under similar test conditions in our laboratories. A brief summary of the results from these preliminary studies is shown in Table 33.

Satellite groups. In the 16 groups of the satellite study the rats were simultaneously exposed to various combinations of chemicals, all at the MOAEL level. The 16 groups jointly comprise a two-level factorial design with nine factors. Thus, the experiment was a 1/32 fraction of a complete two-level study (i.e. $1/32 \times 2^9$). Each compound is absent in eight of the experimental groups and present in the other eight. For any pair of compounds four of the 16 groups contain both compounds, four groups contain neither, four groups contain only the first one of the pair, and four of the groups contain only the second compound of the pair (see Table 34). The combinations of chemicals in the satellite study were chosen such that the results would allow analysis of the interactions between the

Table 33. Exposure levels and summary of toxic effects of nine chemicals as observed in preliminary studies with the individual compounds in male rats after oral (mg/kg diet) or inhalatory exposure (ppm) for 4 weeks[a]

	MOAEL	NOAEL	1/3 NOAEL	target system at MOAEL	target system above MOAEL
Aspirin	5000	1000	330	BW↓, WBC↑, Throm↓, TP↓, PalmCoA↑, hepatocellular hypertrophy	4xMOAEL: BW↓↓, Hb↓, PCV↓, THROM↓, ASAT-↑↑↑, ALAT↑↑, ALP↑, CREAT↓, TP↓↓, PalmCoa↑↑, Hepato-cellular hypertrophy, Hyperkeratosis in stomach, RW liver ↑↑, RW kidney ↑↑, RW Spleen↑↑
$CdCl_2$	50	10	3	FC↓, Hb↓, PCV↓, MCV/ MCH↓, ALAT↑	2xMOAEL: BW↓, Hb↓, PCV↓, MCV/MCH↓, ALAT/ASAT↑↑
$SnCl_2$	3000	800	260	BW↓, Hb↓, ALP↓	3.3xMOAEL: BW↓↓, HB↓, ALP↓
Loperamide	30	6	2	FC↓, WBC↓	3.3xMOAEL: FC↓, THROM↓, GLUC↓
Spermine	2000	400	130	K↓, Myocardial degeneration	2.5xMOAEL: BW↓↓, FC↓↓, Hb↓, PCV↓, THROM↑. MCV/MCH↓, ASAT/ALAT↑↑↑, CREAT↑, GLUC↑↑, Inorg-P↑↑, Ca↓↓, Myocardial degeneration, Hepato-cellular hypertrophy, Spermatogenesis, RW heart↓, RW liver↓, RW spleen↑↑, RW heart↑↑
BHA	3000	1000	330	Hyperkeratosis oesophagus	3.3xMOAEL: BW↓, Hyperkeratosis of oesophagus and stomach
DEHP	1000	200	65	PalmCoa↑,	5xMOAEL: BW↓, Hb↓, PCV↓, PalmCoa↑↑↑, TP↓, hepatocellular hypertrophy, RW liver↑↑
Dichloromethane	500	100	30	CO-HB↑↑,	5xMOAEL: CO-Hb↑↑, HB↓, PCV↓, THROM↑
Formaldehyde	3	1	0.3	Hyperplasia respiratory epithelium↑↑	3.3xMOAEL: BW↓, hyperplasia of respiratory epi-thelium ↑↑↑

[a]Only the parameters showing treatment-related changes are included in this Table. Studies were performed with groups of 10 animals (for Cd, DEHP, MC and aspirin with groups of 8 animals. For the parameters PalmCoa, ASAT, ALAT and ALP 1, 2 or 3 arrows indicate a change of 100%, more than 100%, more than 200% compared to control. For the parameter BW, FC, Gluc, RWorgan one arrow means up to 20% change, two arrows means 20-50% change. For all other parameters one arrow indicates up to 10% change, and two arrows up to 20%. BHA=Butyl hydroxyanisol, DEHP=Di(2-ethylhexyl)phthalate.

Table 34. Test groups and exposure levels of a 4-week toxicity study with combinations of nine compounds in male rats.

	Formaldehyde (For)	Dichloromethane (MC)	Aspirin (Asp)	$CdCl_2$ (Cd)	$SnCl_2$ (Sn)	Loperamide (Lop)	Spermine (Sper)	BHA	DEHP
				Main groups[a]					
4 Groups									
Control	-	-	-	-	-	-	-	-	-
1/3 NOAEL[b]	+	+	+	+	+	+	+	+	+
NOAEL[b]	+	+	+	+	+	+	+	+	+
MOAEL[b]	+	+	+	+	+	+	+	+	+
				Satellite groups[a]					
16 Groups[c]									
For	+	-	-	-	-	-	-	-	-
Sn/MC/Lop/Asp	-	+	+	-	+	+	-	-	-
Cd/MC/Sper/Asp	-	+	+	+	-	-	+	-	-
Sn/Cd/Sper/Lop/For	+	-	-	+	+	+	+	-	-
BHA/MC/Sper/Lop	-	+	-	-	-	+	+	+	-
Sn/BHA/Sper/Asp/For	+	-	+	-	+	-	+	+	-
Cd/BHA/Lop/Asp/For	+	-	+	+	-	+	-	+	-
Sn/Cd/BHA/MC	-	+	-	+	+	-	-	+	-
DEHP/Sper/Lop/Asp	-	-	+	-	-	+	+	-	+
Sn/DEHP/MC/Sper/For	+	+	-	-	+	-	+	-	+
Cd/DEHP/MC/Lop/For	+	+	-	+	-	+	-	-	+
Sn/Cd/DEHP/Asp	-	-	+	+	+	-	-	-	+
BHA/DEHP/MC/Asp/For	+	+	+	-	-	-	-	+	+
Sn/BHA/DEHP/Lop	-	-	-	-	+	+	-	+	+
Cd/BHA/DEHP/Sper	-	-	-	+	-	-	+	+	+
MOAEL[d]	+	+	+	+	+	+	+	+	+

[a]Main groups and satellite groups comprised 8 and 5 animals respectively. [b]Dose levels of MOAEL and NOAEL are given in Table 33. [c]In the satellite study all compounds were dosed at the MOAEL. [d]Data of 5 animals of the main study were used as the 16th group of the satellite study.

nine chemicals (two-factor interactions), but would also allow for an optimal analysis between the main effects of the individual compounds (see Statistical Analysis).

Except for the 'all MOAEL group', the satellite study had five animals per group. The animals were allocated to the groups by restricted randomisation, using the quintile of their body weight as allocation criterion. Thus, before randomisation there were five initial body-weight groups. Each treatment group had one randomly chosen animal from each of the body-weight groups. The main study, including the all-MOAEL group had eight animals per group. By analogy with the satellite study, there were eight initial body-weight groups. This randomisation was carried out separate from the randomisation of the satellite study. When complementing results of the 15 satellite groups with the all-MOAEL group, we used the animals from the five main-study body weight groups with a mean closest to the five satellite-study body weight groups. In subsequent analysis, we used a blocking variable to model the difference between the body-weight groups. We were apparently successful in that, because the grouping accounted for a substantial part of the variation in many parameters.

Test Diets

Except for formaldehyde and dichloromethane all test substances were administered *ad libitum* via the diet, at constant dietary concentrations, for 4 weeks. Test diets were prepared by blending the test compounds and cereal based rodent diet in a Stephan cutter, and were then stored in sealed plastic bags in a freezer at -20 °C. Twice a week the feeders were refreshed with the test diets. Analysis of the diets revealed that the actual concentration of the test compounds at 1/3 NOAEL, NOAEL and MOAEL was as follows (expressed as a percentage of intended value): 75, 84 and 96% for aspirin, 95, 96 and 95% for BHA, 90, 86 and 100 % for cadmium chloride, 94, 87 and 88 % for DEHP, 76, 91 and 102% for spermine and 92, 82 and 83 % for stannous chloride. Loperamide was only analysed at the MOAEL level and concentration was 103% of the intended level. All compounds were stable during a 14 day storage period.

Inhalatory exposure

Animals were exposed to formaldehyde and dichloromethane by inhalation in H-1000 multitiered inhalation chambers manufactured by Hazleton Systems Inc., USA. The chambers are illuminated externally by normal laboratory TL-lighting. The number of air changes was at least 12 per hour. The rats were exposed for 6 h a day, five days a week, during four weeks resulting in a total number of 20 exposure days. Dichloromethane was evaporated by bubbling pressurised air through the test material. Formaldehyde gas was generated by dissolving paraformaldehyde in water, and vaporisation of this solution under heating. The generated mixture of air with formaldehyde and/or dichloromethane was diluted with filtered air from the air-conditioning system to obtain the desired test concentrations. The concentration of dichloromethane in the test atmosphere was determined by total carbon analysis using a flame ionisation detector (Carlo Erba, Italy). Formaldehyde concentrations were determined colorimetrically by means of the Hantzsch reaction using an analysing system of Skalar Analytical (Breda, Netherlands). Atmosphere samples were taken automatically from each of the chambers in an alternating order at a location close to the animals' site. The sampling lines were heated to avoid condensation of the test material during sampling and transportation. Each chamber was monitored about once every hour. Analysis of the test atmosphere showed that the actual concentrations of dichloromethane were 36±2.1, 103±5.0, 520±20 ppm for 1/3 NOAEL, NOAEL and MOAEL respectively. The mean actual concentrations of formaldehyde during the 20 exposure days were 0.35±0.03 ppm, 1.09±0.1 ppm and 3.1±0.25 ppm for 1/3 NOAEL, NOAEL and MOAEL, respectively.

General observations

General health status of the rats was recorded twice a day. Body weights (BW) were recorded individually on the first day of the study and weekly thereafter. Food consumption (FC) was measured once every week by weighing the feeders.

Haematology

On day 23 and 24 blood samples were taken from the tip of the tail of all animals and examined for haemoglobin (Hb) concentration, packed cell volume (PCV), red blood cell count (RBC), white blood cell count (WBC), and thrombocytes (THROM) using a Sysmex K-1000 Haematology Analyzer (Toa Medical Electronics Co. Ltd, Japan). Prothrombin time (PTT) was determined using a Normotest kit. Methaemoglobine (MHb) and carboxy haemoglobin (CO-Hb) concentration were determined within 1 h after sampling according to the method of Brown (1980) and Bauer (1974) respectively. The mean corpuscular volume (MCV), mean corpuscular haemoglobin (MCH) and mean corpuscular haemoglobin concentration (MCHC) were calculated.

Clinical Chemistry

Clinical chemistry was conducted at autopsy in blood collected from the abdominal aorta of all rats. Blood was collected in heparinised plastic tubes and centrifuged at 1250 *g* for about 15 minutes, using Sure-sep II dispensers from General Diagnostics and than analysed for alkaline phosphatase (ALP), alanine aminotransferase (ALAT), aspartate aminotransferase (ASAT), gamma glutamyl transpeptidase (GGT), total protein (TP), cholesterol (CHOL), triglycerides (TriGlyc), albumin (ALB), urea, creatinine (CREAT), total bilirubin (Bili-Tot), non-fasting glucose (GLUC), sodium, potassium, calcium, chloride and inorganic phosphate. Analyses were performed on a Cobas-Bio centrifugal analyser using Baker and Boehringer reagent kits. Sodium and Potassium were analysed using a Electrolyte-2-analyzer, chloride was analysed with a Chloro counter.

Biochemistry

At autopsy the left lateral lobe of the liver was preserved in liquid nitrogen. The tissue was thawed and homogenized using a teflon pestle in a 10mM Tris-HCl buffer (pH 7.4, 4 °C). The palmitoyl-CoA oxidase activity (PalmCoA) was determined in the total liver homogenates after a triple thaw/freeze procedure according to Reubsaet *et al* (1988).

Pathology

On day 28 rats were killed by exsanguination via the abdominal aorta under ether anaesthesia and then examined grossly for pathological changes. The weights of the adrenals, heart, kidneys, spleen, testes, lungs and liver were recorded and the ratios of organ weight to body weight (i.e. relative weight, RW) were calculated. These organs, and in addition oesophagus and stomach were preserved in neutral aqueous phosphate-buffered formalin solution, embedded in paraffin wax, sectioned at 5 µm, stained with haematoxylin and eosin and examined microscopically. The head was removed from each of the carcasses and was flushed retrograde through the nasopharyngeal orifice with 10 ml 4% buffered formaldehyde solution. Thereafter the head were decalcified and sections of the nose (5 µm) were prepared at described previously (Cassee *et al*., 1996b). Nose section levels 2 and 3 were examined for histopathological changes. In the main part of the study microscopic examination was carried out of all organs, in the satellite part of the study only target organs (i.e. liver and nasal passages) were examined microscopically.

Statistical analysis

Main groups & Satellite groups. Body weights on day 28 were evaluated by one-way analysis of covariance followed by an application of the multiple comparison test of Dunnett (1964). Organ weights, haematology, clinical chemistry and biochemistry values were evaluated by one-way analysis of (co)variance followed by an application of the Dunnett's multiple comparison test. Histopathological data were evaluated by an exact test based on the hypergeometric distribution. This is commonly called Fisher's exact test.

Satellite groups. If all possible combinations of the compounds would have been tested, the study would have comprised 2^9 (=512) combinations of chemicals. However, at most 16 satellite groups could be managed experimentally. Therefore the satellite part of the study comprised a 1/32 fraction of a two-level factorial design with nine factors (= nine chemicals) in 16 experimental groups as described by Box *et al*. (1978). The use of a fractional factorial design implicates that we must deal with a complex confounding pattern between main effects and interactions (main effects are aliased with interactions and interactions are aliased with each other). First, the complete aliases pattern in a 1/32 fraction of 2^9 design was worked out in code form (A,B...J) according to Box *et al*. (1978). The next step was to assign chemicals to the codes and formaldehyde was chosen for code J. The confounding pattern of the present study with the aliases between two-factor interactions and main effects is given in Table 35. We intentionally choose formaldehyde aliased with two factor interactions because at the concentration levels chosen formaldehyde acts as a local irritant in the nose and we can assume that there will be no further interaction with systemic effects of the other compounds in the study. This will greatly help the interpretation of the aliased pattern since there will be either responses without any effect of formaldehyde (omit these effects from the list in Table 35) or responses without interactions between two non-formaldehyde effects (omit these effects from the list). Therefore, J=formaldehyde is a good choice to interpret the confounding pattern. To analyze

Table 35. Confounding pattern between compounds in the satellite groups using a fractionated 2^9 design with 9 chemicals in 16 experimental groups in a 4-week toxicity study[a].

Aliasing between main effects and two-factor interactions
Sn=Sper.Lop
Cd=Lop.For
BHA=Asp.For
DEHP=Cd.For
MC=DEHP.For
Sper=Sn.For
Lop=Cd.For
Asp=BHA.For
For=DEHP.MC=Sn.Sper=Cd.Lop=BHA.Asp
Aliasing between two-factor-interactions
Sn.Cd=BHA.MC=Sper.Lop=DEHP.Asp
Sn.BHA=Cd.MC=DEHP.Lop=Sper.Asp
Sn.DEHP=MC.Sper=BHA.Lop=Cd.Asp
Sn.MC=Cd.BHA=DEHP.Sper=Lop.Asp
Sn.Lop=BHA.DEHP=Cd.Sper=MC.Asp
Sn.Asp=Cd.DEHP=BHA.Sper=MC.Lop

[a]Test groups and complete aliasing pattern in a 1/32 fraction of 2^9 design were first worked out in coded form (A,B...J; Box *et al*., 1978). Next chemicals were assigned to the codes. J was allotted to For. Aliases between two- and three-factor interactions are not evaluated. Abbreviations of chemical names are shown in Table 34.

the aliasing of first-order interactions with each other: when such a 'string' is statistically significant, we know that at least one of the interactions is active. Which one? (1) determine the active main effects, (2) determine whether there is an interaction in the string involving two active main effects. If there is (are): this (these) are the most probable ones. If not: (3) determine whether there are interactions involving one of the active main effects. In any case, use the opinion of the experts.

If all compounds would show effect-addition, the effect of any compound could be calculated by subtracting the mean of the groups not containing the compound from the mean of the other groups. The general balance in the design ensures that the effects of the other compounds are cancelled out. Thus the joint effect of the compounds could be decomposed into nine main effects; see also the example of the General introduction.

It is plausible that some compounds will not exhibit effect-addition and in these cases the term non-additivity is applied to describe the interaction between the compounds. A measure of non-additivity is the difference between the effect of a compound in the presence of another one and the effect of the compound in the other one's absence. It follows from the properties of the experimental design that the measures of non-additivity are confounded (i.e. aliased) in seven sets of four confounded first-order interactive effects. In one of these sets the interactive effects are also confounded with the main-effect of formaldehyde. Any interaction of formaldehyde with another compound is confounded with a main-effect not due to the other compound. We assume that cases of interaction (read non-additivity) of two compounds did not depend on the presence of a third one. Thus, three factor interactions are not taken into account.

Statistical significance of the effects concerning body weights, organ weights, haematology, clinical chemistry and biochemistry values, respectively, was evaluated with *t*-tests according to Fisher (1935). In this approach multiple t-test are performed only when a preliminary *F*-test turns out to be statistically significant (5 % level). This method is preferred to the popular Bonferroni method. The latter method dictates performing single tests at a level of 0.0033 (=0.05/15) to obtain a joint significance level of the 15 tests carried ot on one and the same response of at most 0.05. Through its restrictive level of significance, the Bonferroni method is less powerful than the Fisher method (Hochberg and Tamhane, 1987).

There were 15 independent tests for each parameter, each test corresponding to one of the effects as given in Table 35. Statistical significance in a group of confounded compounds implies statistical significance of main effects or two-factor interactions. Which one(s) is a matter of expert-interpretation as described above, or of further research. Since many effects are specific for one or two of the compounds, expert-interpretation in this sense will hardly be necessary. Thus we were able to attribute most cases of statistical significance of confounded groups as shown in Table 35 to only one of the several possibilities. (i.e. main effect or two-factor interaction(s)).

For the analysis of the carboxy-Hb findings, the data were reduced to 0 = absent and 1 = present. These data, as well as the histopathological findings, were evaluated with generalised linear models that use the logit-link and a binomial error-distribution (McCullagh and Nelder, 1989). For the satellite groups, we used a forward selection procedure and initially tried main effects only. Main effects were included in the model whenever the corresponding decrease in deviance exceeded 3.84. In a second step, interactions were fitted as well. For the reduced data of the main study the following independent between-groups comparisons were studied; 1) control versus 3 dosage groups, 2) 1/3 NOAEL versus NOAEL, 3) NOAEL versus MOAEL.

Results

All rats survived to scheduled autopsy. No relevant clinical signs were observed during the study. The haematological parameters PTT, RBC, and MCHC were not affected. The clinical

chemistry of the plasma did not show relevant changes in inorg-P, citrate, sodium, urea, and chloride. These parameters are therefore not presented in the tables.

Main groups

Compared to the control group, only rats in the MOAEL group showed growth retardation and a decrease in food intake (data of day 28 are shown in Table 36). Combined exposure of the nine compounds at the MOAEL resulted in a significant decrease in MCV and thrombocyte counts (for haematology parameters see Table 36). CO-Hb concentration was significantly increased in the NOAEL and MOAEL groups (3.7 and 11.0 % respectively) and values were similar to those found in the preliminary study in rats that were exposed to dichloromethane alone (cf. Table 33). A CO-Hb formation of less than 5 % is not considered to be of toxicological significance (WHO, 1987). There were no other haematological effects in the main groups.

At the MOAEL, many effects on clinical chemistry (Table 37) were encountered such as a decrease in ALP activity, a decrease of the glucose, triglycerides and cholesterol concentrations, and an increase in ALAT/ASAT activities, albumin, bilirubin and total protein concentrations. In the lower dose groups only a few statistically significant changes in clinical chemistry were noted: a decrease in ALP activity and triglyceride concentration in the rats of the NOAEL group, and a decrease in bilirubin concentration in the rats of the 1/3 NOAEL group. The change in bilirubin levels was not seen in the preliminary studies and there was not a clear dose-effect relationship. Therefore, it is doubtful whether this bilirubin effect can be attributed to the treatment.

In the MOAEL group the weights of heart, and spleen were decreased and the liver weight was increased (Table 38). The relative weights of all organs (except those of spleen and heart) were significantly higher in rats of the MOAEL group (Data not shown). Relative kidney weights were also increased in rats of the NOAEL and 1/3 NOAEL groups (Table 38).

Gross examination at autopsy was essentially negative. There were no treatment-related histopathological changes in the adrenals, heart, oesophagus, spleen, testes, lungs or larynx. Treatment-related histopathological changes were seen in the liver, and nasal cavity in the MOAEL group, but also in the NOAEL group (Table 39). In the livers of all rats of the MOAEL group hepatocellular hypertrophy was observed. Similar, but much less severe and less frequent changes in the liver were seen in all rats of the NOAEL group. In the range-finding study histopathological liver changes had been observed in rats exposed to DEHP or aspirin at the MOAEL, and occurred also in rats exposed to spermine at levels above the MOAEL (cf. Table 33).

Table 36. Mean body weights, food consumption and haematological findings in male rats exposed for 4 weeks to selected combinations of maximally nine compounds via the food or by inhalation[a]

	BW28 g	FC28 g/rat/day	PCV l/l	HB mmol/l	MCV l/l	MCH fmol	CO-Hb %	MHb %	WBC 10^9/l	THROM 10^9/l
					Main groups					
Control	304.3±21.1	20.1±0.4	0.439+0.015	9.8±1.2	50.9±0.4	1.13±0.05	0.8±1.8	1.16±0.21	12.6±2.6	597±31
1/3 NOAEL	302.5±14.3	19.9±0.2	0.434+0.005	9.9±0.3	50.7±0.6	1.15±0.01	1.1±1.0	1.11±0.16	11.3±1.5	561±72
NOAEL	297.6±20.7	19.3±0.8	0.425+0.009	9.7±0.6	50.5±0.5	1.15±0.02	3.7±1.5**	1.29±0.22	11.5±1.3	549±64
MOAEL	262.8±20.4**	18.2±0.6	0.446+0.006	10.2±0.3	50.1±0.5*	1.14±0.01	11.0±1.3**	1.35±0.50	13.1±2.6	457±135**
					Satellite groups					
For	304.5±25.5	19.0±1.1	0.453+0.009	10.2±0.3	50.7±0.3	1.15±0.02	0.0±0.0	1.64±0.38	11.5±1.4	571±28
Sn/MC/Lop/Asp	269.1±17.5	18.4±0.4	0.460+0.008	10.5±0.2	51.6±0.3	1.18±0.03**	8.0±1.5**	1.33±0.27	12.0±1.9	438±95 **
Cd/MC/Sper/Asp	267.4±20.8	18.7±0.5	0.427+0.008	9.8±0.3	48.8±1.4**	1.12±0.03	9.3±0.9**	1.22±0.25	11.0±1.8	465±35*
Sn/Cd/Sper/Lop/For	285.0±21.1	19.1±0.6	0.427+0.011	9.8±0.5	49.9±0.4	1.14±0.02	0.0±0.0	1.44±0.59	10.8±1.0	537±102
BHA/MC/Sper/Lop	286.3±14.2	19.6±0.3	0.450±0.010	10.4±0.5	50.9±0.3	1.18±0.01**	2.7±2.2	1.35±0.21	9.9±1.1*	526±113
Sn/BHA/Sper/Asp/For	270.0±9.7	19.3±0.5	0.433±0.004	9.9±0.2	50.2±0.9	1.15±0.02	0.0±0.0	1.45±0.15	9.1±1.0*	478±63
Cd/BHA/Lop/Asp/For	266.6±14.4	18.9±0.3	0.427±0.003	9.7±0.2	49.0±0.8**	1.12±0.01	0.0±0.0	1.38±0.34	11.4±1.3	531±34
Sn/Cd/BHA/MC	283.6±19.2	19.6±0.6	0.414±0.006	9.5±0.3	48.7±0.4**	1.12±0.03	7.1±2.3**	1.15±0.41	8.3±0.4**	592±40
DEHP/Sper/Lop/Asp	270.7±20.2	17.9±0.5	0.441±0.007	10.1±0.2	51.4±0.5	1.18±0.01**	0.7±1.0	1.31±0.20	9.5±3.1*	474±70*
Sn/DEHP/MC/Sper/For	290.0±16.8	18.9±0.4	0.442±0.004	10.1±0.3	51.2±0.8	1.17±0.02*	6.4±2.5**	1.32±0.17	10.2±0.7	552±56
Cd/DEHP/MC/Lop/For	287.2±16.0	18.6±0.4	0.418±0.010	9.6±0.5	48.0±1.0**	1.10±0.03	3.9±2.7	1.23±0.25	11.0±0.7	593±77
Sn/Cd/DEHP/Asp	264.5±13.9	18.1±0.4	0.436±0.009	10.1±0.3	49.8±0.3	1.16±0.03	0.0±0.0	1.34±0.43	10.5±1.4	497±57
BHA/DEHP/MC/Asp/For	282.1±19.1	19.3±0.4	0.447±0.006	10.3±0.3	51.6±0.4	1.19±0.02**	3.3±0.5	1.33±0.46	9.5±0.9*	543±41
Sn/BHA/DEHP/Lop	289.3±13.0	19.8±0.5	0.422±0.003	9.7±0.1	50.1±0.6	1.15±0.02	0.0±0.0	1.26±0.32	10.4±1.8	590±19
Cd/BHA/DEHP/Sper	289.5±22.3	19.2±0.5	0.416±0.004	9.5±0.2	48.0±0.6**	1.10±0.01	0.0±0.0	1.32±0.37	9.6±1.6*	596±70

[a]Parameters showing treatment-related changes. Values are the means ± s.d. for groups of 8 animals (main groups) or 5 animals (satellite groups). BW28: body weight on day 28. FC28: means over 4 weeks. The values marked with asterisks differ significantly from the controls. Tests were as follows. Body weight: analysis of covariance and Dunnett's multiple comparisons test (body weight on day 0 as covariate); haematology parameters excepting CO-Hb and MetHb: ANOVA and Dunnett's multiple comparisons test; CO-Hb and MetHb: Kruskal-Wallis and Mann-Whitney U-tests. Multiple comparisons tests were two-sided. * $P<0.05$; ** $P<0.01$.

Table 37. Mean values of clinical chemistry in plasma and biochemistry in liver of male rats exposed for 4 weeks to selected combinations of maximally nine compounds via the food or by inhalation

	GLUC mmol/l	ALP U/l	ALAT U/l	ASAT U/l	TP g/l	ALB g/l	Bili-Tot µmol/l	CHOL mmol/l	Triglyc mmol/l	CREAT µmol/l	PalmCoA µmol/l/min
						Main groups					
Control	9.52±0.83	306±50	47±7	70±7	61±1	32±1	2.0±0.4	1.65±0.10	0.96±0.20	22±1	0.44±0.08
1/3 NOAEL	8.96±0.70	278±16	43±4	65±7	61±1	32±0	1.5±0.3**	1.64±0.11	0.89±0.23	20±1	0.45±0.07
NOAEL	9.47±0.39	254±45*	47±8	71±10	60±1	32±1	1.8±0.2	1.62±0.09	0.68±0.20*	22±1	0.53±0.09
MOAEL	8.28±0.32**	227±21**	59±6**	82±7*	58±2**	34±1**	2.5±0.4*	1.40±0.11**	0.43±0.14**	23±1*	1.60±0.52**
						Satellite groups					
For	7.94±0.42	281±26	46±4	70±7	61±2	32±1	1.7±0.2	1.82±0.19	0.90±0.16	21±1	0.34±0.02
Sn/MC/Lop/Asp	8.03±0.92*	234±17**	47±9	71±6	56±1**	32±1	2.3±0.1	1.36±0.12**	0.54±0.11**	25±1	0.88±0.23*
Cd/MC/Sper/Asp	8.01±0.87	268±34*	81±11**	86±13	57±1**	33±1	2.8±0.2	1.28±0.10**	0.50±0.12**	24±1	0.73±0.34
Sn/Cd/Sper/Lop/For	8.49±0.48	180±6*	53±7	75±7	59±3	32±1	1.8±0.3	1.64±0.09	0.62±0.05**	20±1	0.39±0.10
BHA/MC/Sper/Lop	8.48±0.60	226±35**	48±8	65±3	62±2	33±1	1.8±0.3	1.78±0.10	0.65±0.12**	20±1	0.38±0.06
Sn/BHA/Sper/Asp/For	7.71±0.88**	210±17**	44±7	70±11	58±1	33±1	2.0±0.3	1.53±0.09	0.59±0.15**	23±1	0.64±0.25
Cd/BHA/Lop/Asp/For	8.61±1.26	221±10**	88±7**	96±11*	58±2	33±1	2.0±0.4	1.59±0.06	0.48±0.11**	24±1	0.75±0.16
Sn/Cd/BHA/MC	8.09±0.44	174±9**	46±4	65±7	60±2	33±1	1.8±0.2	1.86±0.06	0.64±0.17**	20±1	0.35±0.03
DEHP/Sper/Lop/Asp	9.66±1.05	323±29	52±6	77±12	57±1**	33±1	2.6±0.4	1.29±0.08**	0.47±0.12**	25±1	1.35±0.45**
Sn/DEHP/MC/Sper/For	8.22±0.63	222±11**	43±17	71±16	59±1	33±1	1.2±0.2**	1.44±0.07*	0.49±0.20**	20±1	1.22±0.15**
Cd/DEHP/MC/Lop/For	8.47±0.88	244±20***	80±9**	88±14*	59±3	34±1	1.5±0.1	1.40±0.07**	0.43±0.08**	23±1	0.92±0.16*
Sn/Cd/DEHP/Asp	8.48±0.93	236±13**	52±4	80±10	56±1	33±1	2.6±0.3*	1.36±0.06**	0.50±0.13**	25±2	1.44±0.28**
BHA/DEHP/MC/Asp/For	8.44±0.50	246±14**	49±3	68±12	60±2	34±1**	2.0±0.4	1.53±0.17	0.67±0.15**	25±2	1.36±0.32**
Sn/BHA/DEHP/Lop	8.78±0.61	184±7**	44±8	69±8	61±1	34±1	1.2±0.2**	1.91±0.14**	0.58±0.16**	19±2*	0.59±0.17
Cd/BHA/DEHP/Sper	8.60±0.73	204±14**	69±7**	72±7	62±2	34±1**	1.5±0.4	1.75±0.12	0.75±0.25*	19±1*	0.63±0.16

[a]Parameters showing treatment-related changes. Values are the means ± s.d. for groups of 8 animals (main groups) or 5 animals (satellite groups). The values marked with asterisks differ significantly from the controls (ANOVA and Dunnett's multiple comparisons test on log-transformed parameters, two-sided).* P<0.05; ** P<0.01.

Table 38. (Relative) Organ weights of male rats exposed for 4 weeks to selected combinations of maximally nine compounds via the diet or by inhalation

	Wadrenals g	Wkidneys g	RWkidneys g/kg	Wspleen g	Wheart g	Wliver g	Wlung g
			Main groups				
Control	0.049±0.005	2.22+0.19	7.29±0.26	0.534±0.034	1.03±0.03	10.9±0.97	1.26±0.10
1/3 NOAEL	0.052±0.005	2.34+0.11	7.73±0.31*	0.526±0.033	0.98±0.02	10.95±0.81	1.24±0.05
NOAEL	0.052±0.005	2.35+0.14	7.91±0.36**	0.509±0.033	0.95±0.02	11.17±1.05	1.25±0.08
MOAEL	0.048±0.005	2.24+0.17	8.52±0.21**	0.464±0.033**	0.91±0.02**	11.51±1.03*	1.17±0.08
			Satellite groups				
For	0.053±0.003	2.12±0.20	7.37±0.25	0.52±0.05	1.01±0.14	10.46±0.81	1.22±0.05
Sn/MC/Lop/Asp	0.048±0.003	2.24±0.21	8.25±0.16**	0.48±0.02	0.92±0.09	10.18±0.44	1.13±0.11
Cd/MC/Sper/Asp	0.049±0.003	2.12±0.25	7.90±0.26**	0.47±0.02	0.88±0.07	9.85±1.12	1.09±0.08
Sn/Cd/Sper/Lop/For	0.055±0.008	2.17±0.23	7.54±0.15	0.50±0.03	1.01±0.05	9.79±1.0	1.17±0.11
BHA/MC/Sper/Lop	0.051±0.005	2.16±0.11	7.51±0.18	0.48±0.04	0.96±0.05	10.61±0.55	1.22±0.08
Sn/BHA/Sper/Asp/For	0.048±0.003	2.34±0.21	8.62±0.34**	0.48±0.05	0.92±0.05	10.07±1.20	1.16±0.05
Cd/BHA/Lop/Asp/For	0.048±0.005	2.19±0.08	8.17±0.20**	0.46±0.04	0.92±0.03	10.06±0.92	1.16±0.10
Sn/Cd/BHA/MC	0.050±0.005	2.19±0.21	7.68±0.33	0.49±0.05	0.97±0.11	10.68±1.37	1.20±0.11
DEHP/Sper/Lop/Asp	0.048±0.003	2.19±0.31	7.99±0.34**	0.46±0.05	0.94±0.11	10.55±1.23	1.20±0.14
Sn/DEHP/MC/Sper/For	0.051±0.008	2.20±0.20	7.44±0.21	0.50±0.05	0.98±0.08	11.23±1.40	1.18±0.08
Cd/DEHP/MC/Lop/For	0.050±0.003	2.20±0.25	7.24±0.16	0.49±0.03	0.95±0.14	10.84±0.98	1.19±0.11
Sn/Cd/DEHP/Asp	0.047±0.008	2.26±0.15	8.10±0.16**	0.46±0.05	0.87±0.08	10.43±0.70	1.13±0.08
BHA/DEHP/MC/Asp/For	0.048±0.005	2.29±0.22	8.07±0.43**	0.48±0.05	0.98±0.14	12.36±0.67*	1.23±0.05
Sn/BHA/DEHP/Lop	0.055±0.005	2.25±0.14	7.69±0.10	0.50±0.05	0.97±0.08	11.88±0.70	1.22±0.11
Cd/BHA/DEHP/Sper	0.051±0.005	2.25±0.25	7.72±0.27*	0.48±0.05	0.97±0.11	12.13±1.04*	1.19±0.11

[a]Parameters showing treatment-related changes. Values are the means ± s.d. for groups of 8 animals (main groups) or 5 animals (satellite groups). The values marked with asterisks differ significantly from the controls (ANOVA and Dunnett's multiple comparisons test on log-transformed parameters, two-sided).* $P<0.05$; ** $P<0.01$.

The nasal cavity of all rats in the MOAEL group showed very slight to moderate hyperplasia of the respiratory and transitional epithelium and/or very slight to moderate squamous metaplasia of both epithelia at cross-level 2. Surprisingly, similar nasal changes were seen in all rats of the NOAEL group, although to a lesser degree (Table 39). From the results of the preliminary studies with the individual compounds it was clear that such effects on the nasal epithelium were to be attributed to formaldehyde, and not to any of the other compounds. However, it was also known that rats exposed to 1 ppm formaldehyde alone would not show such or any other nasal effects (Woutersen *et al*., 1987).
A few rats showed focal inflammatory cell infiltrates and/or alveolar haemorrhages in the lungs (data not shown). The incidence of these lesions were slightly (though not statistically significantly) higher in the animals of the MOAEL group than in the controls.

Satellite groups

In the satellite study, the rats were exposed at the MOAEL only; compared to the controls,

Table 39. Type and incidence of histopathological changes in the nasal cavity (level 2) and liver of rats exposed for 4 weeks to combinations of nine chemicals via the diet or by inhalation[a]

	Control	1/3NOAEL	NOAEL	MOAEL
Nasal cavity				
Hyperplasia of respiratory epithelium				
very slight	0	0	0	0
slight	0	1	0	4
moderate	0	0	1	4
total	0	1	1	8*
Hyperplasia of transitional epithelium				
very slight	1	0	4	0
slight	0	2	2	4
moderate	0	0	2	4
total	1	2	8*	8*
Squamous metaplasia of respiratory/transitional epithelium				
very slight	0	2	1	3
slight	0	0	0	3
moderate	0	0	2	1
total	0	2	3	7*
Inflammatory cell infiltration				
very slight	1	1	0	1
slight	3	3	1	3
moderate	0	0	1	1
total	4	4	2	5
Liver				
Hepatocellular hypertrophy				
very slight	0	0	3	4
slight	0	0	1	4
total	0	0	4	8*
Aggregates of RES cells and necrotic hepatocytes				
slight	0	1	0	0
Perivascular mononuclear-cell infiltrate				
slight	0	0	2	1

[a]8 animals were examined per group. No other treatment-related changes in any of the examined organs were found as compared to controls (Fisher's Exact test, * $p<0.05$).

several changes in haematological and clinical data (Tables 4-6) were encountered. To analyse main effects of individual compounds and interactions between the compounds the data set as shown in the Tables 4-6 was subjected to a factorial analysis as described under Material and Methods. With the application of the factorial design it was possible to identify main effects of the individual compounds and interaction (cases of non-additivity) between two compounds.

A final equation to describe the value of the parameter in any particular mixture in terms of the variables (i.e. compounds) tested was defined (Box, *et al*., 1978) as:

$$Variable_{mix} = Mean + ½*effect_A * A + ½*effect_{AB} * A * B \text{ etc.}$$

where "$Variable_{mix}$" is the total value for the variable in any particular mixture chosen, "Mean" is the overall mean from 16 experimental groups, "$effect_A$" is the mean effect of compound A, B, etc., "$effect_{AB}$" is the interactive effect between compound A and B, and where "A" and "B" have a value of either +1 or -1, i.e. presence or absence of compound A and compound B. The final equations for the relevant haematological and clinical parameters are shown in Table 40. We will exemplify the analysis of Table 40 by way of the parameters PalmCoA and ASAT.

With respect to the PalmCoA activity in the liver, the factorial analysis revealed that two compounds (aspirin and DEHP) were able to induce PalmCoa activity, whereas BHA slightly reduced the activity. Moreover, there was a rather slight and unexpected, but significant ($P < 0.01$) interaction between BHA and DEHP, which resulted in a decreased total PalmCoA activity. Fig. 10a illustrates the interactive effect of BHA and DEHP and shows how this interaction can be visualised from the two by two plot of the effect of the individual compounds. In the figure, the interactive effect between two compounds is indicated by the absence of parallel lines. In summary, the factorial analysis resulted in the following equation:

$$PalmCoA_{mix} = 0.85 + 0.4*DEHP + 0.3*Asp - 0.1*BHA - 0.07*BHA*DEHP.$$

Regarding the ASAT activity, the analysis resulted in the following equation:

$$ASAT_{mix} = 75.4 - 2.49*Sn + 5.32*Cd + 2.79*Lop + 3.66*Asp + 2.37*Cd*Lop - 2.47*Sn*Cd.$$

This means that there were three compounds able to increase the ASAT activity (cadmium chloride, aspirin, or loperamide). Thus, rats exposed to one of these three chemicals showed a increased ASAT compared to the rats which were not exposed to these chemicals (rats in 8 out of 16 groups). One compound (stannous chloride) was able to decrease the ASAT activity. Two cases of significant interactions were identified, namely the interaction between cadmium chloride and loperamide (Cd*Lop) and between stannous chloride and cadmium chloride (Sn*Cd). In the case of the Cd*Lop interaction, the effect of the combination of the two compounds was 2.37 units/l higher than could be expected on the basis of summation of the effects of the two single compounds cadmium chloride and loperamide (i.e. Cd*Lop = + 2.37). In the case of the Sn*Cd interaction, the effect of the combination of the two compounds was 2.47 units/l smaller than expected on the basis of additivity of stannous chloride and cadmium chloride (i.e. Sn*Cd = - 2.47). Fig. 10b and Fig. 10c exemplify how the interactive effects between Cd and Lop and between Cd and Sn should be interpreted.

Again, the interaction is indicated by the absence of parallel lines between the effect of the two compounds.

For every random selection of mixtures from the 9 compounds tested it is possible to predict the overall effect for any particular parameter with the final equations as shown in Table 40. For instance, for animals exposed to cadmium chloride (Viz. Sn absent, Cd present and Asp absent) the body weight on day 28 can be estimated to be 279.2 -2.4*-1 -3.4.*1 -10.0*-1=288.2 g.

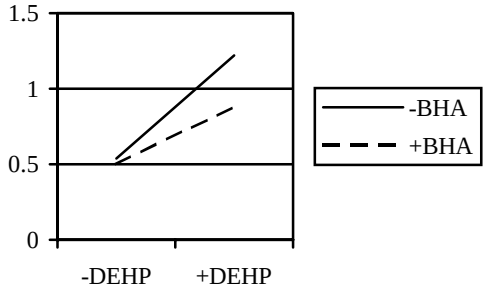
In the case of the parameters ALP and PCV the final equation should be regarded with caution due to the fact that numerous (possible) aliases are present between the identified two-factor interactions. Due to this large number of interactions we can not interpret the

Table 40. Final equations for body weights, food intake and selected clinical and haematological parameters of the satellite groups in terms of main effects and two factor interactions[a]

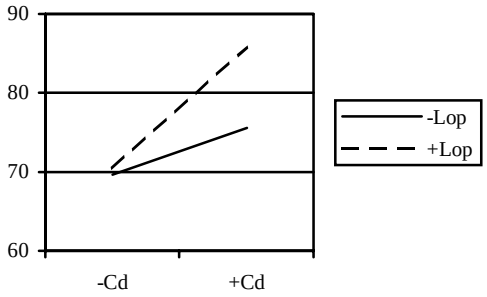
BW28=	279.23 - 2.37*Sn - 3.44*Cd - 10.01*Asp
FI28=	18.89 + 0.31*BHA - 0.33*Asp
WBC=	10.39- 0.42*BHA + 0.45*Lop + 0.43*Lop*BHA + 0.43*Sn*Lop
HB=	9.95 - 0.17*Cd + 0.08*Asp + 0.11*Sn*Cd - 0.14*Cd*MC
PCV=	434.54 - 8.84*Cd - 3.56*BHA + 4.14*Asp + 3.99*Sn*Cd - 3.74*Sn*BHA + 3.57*Cd*DEHP
MCV=	50.0 + 0.19*Sn - 0.98*Cd - 0.17*BHA + 0.32*Asp + 0.40*Sn*Cd + 0.26*Sn*BHA
MCH=	1.15 + 0.005*Sn - 0.02*Cd + 0.005*MC + 0.005*Asp - 0.01* Sn*Cd - 0.005 * Cd*MC
ALB=	33.19 + 0.36*BHA + 0.61*DEHP
ALP=	229.75- 19.84*Sn - 8.71*Cd - 16.75*BHA + 5.54*DEHP + 13.50*Asp + 7.63*Sn*Cd + 9.5*Sn*BHA + 7.79*Cd*BHA + 5.79*Sn*DEHP + 4.84*Cd*DEHP
Bili-Tot=	1.95 + 0.09*Cd - 0.11*BHA + 0.07*Sper + 0.39*Asp - 0.12*BHA*Asp + 0.14*DEHP*Asp
CREAT=	22.0 - 0.34*BHA + 0.32*MC + 1.96*Asp
CHOL=	1.56 - 0.025*Cd + 0.11*BHA - 0.05*DEHP - 0.06*MC - 0.045*Sper - 0.15*Asp + 0.025*Sn*Cd + 0.025*BHA*DEHP
GLUC=	8.39 + 0.22*DEHP + 0.21*Lop
THROM=	529.71 - 14.84*Sper - 39.61*Asp
MHb=	1.33 - 0.07*MC
PTT=	40.41 - 0.56*BHA + 0.57*DEHP + 0.65*MC + 0.81*Asp
ASAT=	75.39 - 2.49*Sn + 5.32*Cd + 2.79*Lop + 3.66*Asp + 2.37*Cd*Lop + 2.47*Sn*Cd
ALAT=	56.40 - 7.58*Sn + 9.8*Cd + 2.63*Lop + 2.6*Asp - 5.53*Sn*Cd
TP=	58.95 - 0.45*Sn + 1.05*BHA - 1.43*Asp + 0.35*BHA*DEHP
NA=	72.75 + 0.51*MC + 0.46*Lop + 0.62*Asp - 0.44*Sn*DEHP
Triglyc=	579.87 - 36.88*DEHP - 33.38*MC - 52.13*Lop - 54.38*Asp + 33.38*DEHP*Asp + 47*MC*Asp
PalmCoA=	0.85 + 0.4*DEHP + 0.3*Asp -0.1*BHA -0.07*BHA*DEHP
Wadren=	0.05 - 0.002*Asp + 0.01*BHA*DEHP
Wkidney=	2.21 - 0.03*Cd + 0.03*BHA
RWkidney=	7.9 + 0.1* Sn + 0.02* BHA + 0.3* Asp
Wspleen=	0.48 - 0.11*Asp
Wheart=	0.94 - 0.28*Asp
Wliver=	10.63 - 0.17*Cd + 0.44*BHA + 0.55*DEHP - 0.16*lop + 0.18*BHA*DEHP
Wlung=	1.18 - 0.015*Cd - 0.02*Asp + 0.02*Asp*DEHP

[a]The equation for two compounds A and B in a mixture showing significant main and interactive effects can be described as: $Variable_{mix}$=Mean + ½*$Effect_A$*A+ ½*$Effect_B$*B + ½*$Effect_{AB}$*A*B, where $Variable_{mix}$ is the value of the variable of any particular mixture chosen, Mean is the mean of the 16 experimental groups, $Effect_A$ is the main effect compound A, $Effect_B$ is the main effect compound B, $Effect_{AB}$ is interactive effect between compound A and B, and A, B indicate the presence of compound A and B (1 or -1, i.e. present or absent). The equation decomposes the overall value of the parameter in any particular mixture in terms of main effects and two-factor interactions. Only significant main effects and two-factor interactions are given ($p<0.05$). If the parameter does not show significant interactions, the chemicals behave in an effect-additive way.

a DEHP / BHA	absent	MOAEL	mean
absent	0.54	1.22	0.88
MOAEL	0.50	0.88	0.69
mean	0.52	1.05	0.79



b Cd / Lop	absent	MOAEL	mean
absent	69.65	75.55	72.60
MOAEL	70.50	85.85	78.18
mean	70.07	80.70	75.39



c Cd / Sn	absent	MOAEL	mean
absent	70.10	85.65	77.87
MOAEL	70.05	75.75	72.90
mean	70.07	80.70	75.39

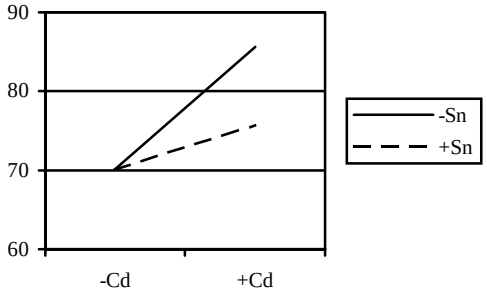


Fig. 10. Two-factor interactions (P<0.05) for the parameters PalmCoa (1a) and ASAT (1b and 1c). Interactions are calculated from the two-by-two-plot of the individual compounds shown on the left. Statistical analysis was performed with the data of the 16 experimental groups of the satellite study. Overall mean effects of the individual compounds can be calculated from the margins of the Tables. Values in the Tables are expressed as µmol/min/ml for PalmCoA activity and as units/l for ASAT activity. The figure on the right illustrates the two-factor interactions: Interactive (=non-additive) effects are indicated by the absence of parallel lines. Details on the analysis are shown in the Materials and Methods section.

confounding pattern and it is rather difficult to identify the main effects of the mixture (cf. Table 35 and 8).

The CO-Hb formation and the histopathological findings, were evaluated in a similar way (see Material and Methods). As expected, the CO-Hb formation was mainly due to the presence of dichloromethane and no interactions with other compounds were observed. There was a significant increase of the incidence of hepatocellular hypertrophy in rats treated with aspirin and DEHP ($P<0.01$). Also, rather unexpectedly, it appeared that rats exposed to cadmium chloride showed a slight increase in the incidence of liver cell hypertrophy ($P<0.05$). In the liver there were no significant cases of non-additivity. In the nose of rats exposed to formaldehyde there was a clear increase in the incidence of total epithelial hyperplasia and squamous metaplasia compared to rats not exposed to formaldehyde ($P<0.01$). Factorial analysis also revealed that in the absence of formaldehyde rats exposed to dichloromethane showed an slightly increased incidence of epithelial hyperplasia ($P<0.05$). In the presence of formaldehyde most of the rats showed epithelial hyperplasia independent of the presence or absence of dichloromethane (data not shown).

Discussion

Exposure at the MOAEL

The main part of the study revealed that combined exposure to nine compounds each at its MOAEL resulted in a wide range of adverse effects. These vary from signs of general toxicity (growth depression, reduced food intake), changes in haematological values (decreased MCV, increased CO-Hb formation, decreased number of thrombocytes), changes in biochemical/clinical parameters (decreased glucose, triglyceride, cholesterol and ALP levels, increased ASAT, ALAT and PalmCoA levels), to increased relative organ weights and histopathological changes in nose, liver and lungs. Based on the toxicity data of the individual compounds, indeed we expected to see most of these effects also in the combination. A few effects seen in the sub-acute toxicity studies with the individual compounds had disappeared in the combination, whereas some effects not seen in the range finding-studies appeared in the combination. For example, the increase in relative weights of kidneys, lungs, testes and adrenals had not been seen in the preliminary studies or at levels higher than the MOAEL of the individual compounds. Although the increase in the relative weight of the adrenals, testes and lungs may be related to the growth retardation (Feron *et al*., 1973, Oichsi *et al*. 1979), the increased liver and kidney weights are considered of toxicological significance. With respect to the kidney weight this view is strongly supported by the dose-related increase in relative kidney weight also seen at the lower dose levels.

The decrease in plasma cholesterol and triglycerides levels in the MOAEL group cannot be explained by the findings with the individual compounds because these parameters had not been included in all of the preliminary studies and were unaffected when they had been measured. Most likely, the effect on cholesterol and triglycerides was caused by the peroxisome proliferator DEHP which has been shown to decrease plasma cholesterol and triglycerides levels in rats (Dirven *et al*., 1990). In the present study, both DEHP and aspirin induced Acyl-CoA oxidase activity (the first enzyme of the peroxisomal β oxidation of fatty acids), and a decreased lipid metabolism seems therefore obvious. Again, we could not verify

this assumption because in the studies with DEHP and aspirin alone, the glycerides and cholesterol plasma contents were not measured. Factorial analysis of the satellite study indeed showed a decreased cholesterol plasma level due to the presence of aspirin and DEHP, although it must be mentioned that rats treated with methylene chloride or spermine also showed slightly decreased cholesterol levels (cf. Table 40).

The interpretation of the results in terms of additive or non-additive effects is hampered, because the overall toxicity of the individual chemicals cannot be ascribed to one specific mechanism of action. For aspirin, loperamide and spermine showing rather aspecific effects in the present study, the interpretation of the data heavily depends on the toxicological end-points measured. For rather characteristic effects such as peroxisome proliferation by DEHP and aspirin, the effect of formaldehyde on the nasal epithelium, CO-Hb formation by dichloromethane, for all of which we assumed dissimilar modes of action, we predicted absence of interactions, and thus response addition. Indeed, in the combination these effects were expressed unaltered, reflecting the concept "independent joint action" (Mumtaz *et al*., 1994).

A thorough analysis to find out whether the adverse effects of the combination of the nine compounds at the MOAEL were less or more pronounced than the effects of the individual compounds, was not carried out on the basis of the results of the main part of the study. The reason is that the results of this part had to be compared with the results of the previous tests on the individual compounds (cf. Table 33) which unavoidably were obtained under slightly different experimental conditions. Moreover, several of the affected parameters had not been included in each of the studies with the individual compounds, which renders a quantitative comparison unreliable. However, in general terms the main part of the study allows the far from spectacular conclusion that simultaneous exposure to chemicals in effective (toxic) doses resulted in effects qualitatively and quantitatively similar or dissimilar to those of the individual compounds. A comparable conclusion has been drawn from our previous studies in rats with combinations of chemicals with different target organs (Jonker *et al*., 1990) or with the same target organ (kidneys), but with different modes of action (Jonker *et al*., 1993).

A more accurate analysis of the interactive effects could be carried out, using the results from the satellite part of the study in which a fractional factorial design (1/32th fraction of a complete design) was applied. Undoubtedly, the confounding pattern would have been less complex and interpretation of the data would have been much easier when a 1/16th fraction or a 1/8th fraction had been applied. However, one has to realise that the number of test groups in such cases will increase to unmanageable numbers, viz. 32 and 64, respectively. For the same reason we restricted the analysis to one dose level, the MOAEL of the individual chemicals, although a three level factorial analysis would have produced relevant information about dose-response curves as has been shown previously (Cassee *et al*., 1996a; Narotsky *et al*., 1995). This also implies that extrapolation of cases of non-additivity to lower dose regions has to be done with great care. If the risk assessor attempts to model the complete dose-response surface of the mixture another approach might be followed in which the interest is focused on departure from additivity of the mixture as a whole, rather than on the identification of specific interactions. This approach suggested by Berenbaum (1985) utilises single chemical dose response information in addition to the observed responses induced by the particular combinations of interest. Clearly, the advantage offered by such an approach becomes apparent upon comparing the number of experimental groups employed to that required by the use of factorial designs (Gennings, 1995). However the fractionated factorial

approach allows more accurate hypotheses as to critical interactions and identifies compounds in the mixture of particular toxicological importance. A drawback of fractionated factorial designs is the possible occurrence of inferences that result from a high level of aliasing, and prior knowledge input needed to be able to interpret the results. In the present study the treatment combinations were selected in a way that allowed the interpretation of the results using the decision process chosen. Thus, to some extent prior knowledge is needed to correctly design this type of studies.

The term 'nonadditivity' was used to describe a statistical interaction between two compounds. The term is not meant to designate an antagonistic or synergistic effect, because without knowledge about the dose-response curves of the individual compounds pure antagonism or synergism cannot be established. It is almost impossible to rule out the possibility that observed interactions are not simple additive effects because full response addition cannot be achieved at dose levels that already have induced maximum or close to maximum responses (binomial or logarithmic dose response curves). For these reasons the interactions for the parameters ASAT and PalmCoa as exemplified in Fig. 10 have to be compared with the dose-response curves established in the preliminary studies. Nevertheless, the fact that statistical interactions were found indicates that the combined effect of two compounds was not a simple summation of effects of the individual compounds. This may be regarded as a warning signal; a careful analysis and maybe further studies of the interactions found are indicated especially in those cases where the combined effect is more pronounced than expected on the basis of full response addition.

Exposure at the NOAEL

Only a few adverse effects were encountered in the NOAEL group: hyperplasia and metaplasia of the nasal epithelium, hepatocellular hypertrophy, decreased plasma triglyceride concentrations, altered ALP enzyme activities and increased (relative) kidney weight. Obviously, combined exposure to compounds at their individual NOAEL can result in a significant effect of the combination which is in line with previous experimental findings (Jonker *et al*., 1990; National Toxicology Program, 1993; Yang, 1994b).

The effect on the nasal epithelium can be attributed to formaldehyde. However, a series of studies has shown that 1 ppm formaldehyde is a NOAEL (Appelman *et al*. 1988, Woutersen *et al*., 1987, Wilmer *et al*., 1987). The finding that 1 ppm formaldehyde in the mixture is not a NOAEL could not be explained. Studies to unravel this remarkable observation are in progress.

Except for a slight increase in the relative kidney weight it was shown that combined exposure below the NOAEL produced no toxic effects. The effect on the kidney weight was also observed in the 1/3 NOAEL group, showing the sensitivity of this rather unspecific toxicological parameter. The satellite part of the study revealed that aspirin, BHA or $SnCl_2$ may increase the kidney weight (Table 40). In the case of aspirin and stannous choride this might due to the decreased body weights. Furthermore there were no cases of potentiating interactions. Therefore, the effect of the mixture of all nine compounds at the MOAEL could be well explained by summation of the effects of BHA, aspirin and $SnCl_2$ (cf. Table 40). A similar joint (additive) action of these compounds might also explain the increase in relative kidney weight at the NOAEL. Comparable effects on kidney weight were also found in the

mixture studies of Jonker *et al*. (1990). A sound explanation for this intriguing and consistent finding is lacking.

In spite of a few observed changes, most end-points were not affected at the NOAEL. Jonker *et al*. (1990), Heindel *et al*. (1994), and Seed *et al*. (1995) reported similar conclusions. Therefore, the most important practical lesson from these studies is that in general combined exposure to the single chemicals does not constitute an evidently increased hazard, provided the exposure level of each chemical is similar to or lower than its own NOAEL.

4.3 Statistically designed experiments in a tiered approach to screen mixtures of *Fusarium* mycotoxins for possible interactions

Introduction

Mycotoxins belong to a group of hazardous compounds that occur simultaneously in food or raw materials. Several classes of these compounds may be produced by a single species of fungus. Particularly hazardous in this respect are the species from the genus *Fusarium*. They can produce Trichothecenes, Zearalenone, and Fumonisins, which have been identified as important contaminants in foodstuffs for human or animal nutrition. Mainly due to the diversity of their chemical structures, these classes elicit a wide range of toxicological effects (Coulombe, 1993).

Several *Fusarium* mycotoxins are often found in combination in infested cereal grains. Because of their natural co-occurrence, there is an increasing concern about the hazard of exposure to mixtures (Doko *et al*., 1996; Tanaka *et al*., 1988). A few studies have been reported that address the toxicity of mixtures of mycotoxins (Bhavanishankar *et al*., 1988; Rotter *et al*., 1992; Diaz *et al*., 1994; Kubena *et al*., 1995), although most of these studies were focused on binary mixtures only. The most obvious reason for the limited number of *in vivo* studies with more complex mixtures of mycotoxins, is the dramatic increase in number of experimental groups with increasing number of compounds in the mixture. Considering the importance of mycotoxin contamination of food there is clear need to develop test strategies to assess the hazard of exposure to more complex mixtures of mycotoxins.

In the quality control of foodstuffs, several *in vitro* bioassays have been proposed to detect mycotoxins (Berger *et al*., 1985; Reubel *et al*., 1987; Buckle and Sanders, 1990; Senter *et al*., 1991). These bioassays are obviously less restrictive in the number of test groups than *in vivo* studies and might therefore be a useful way to start the hazard assessment of complex chemical mixtures. It has been speculated that in some cases the effect of *Fusarium* contaminated samples, as found with the bioassays, could not be explained by the presence of the individual compounds (Porcher *et al*., 1987; Scossa-Romano *et al*., 1987; Norred *et al*., 1990). This might be due to the combined action of individual compounds when present as a mixture in foodstuff. Indeed, for some pairs of Trichothecenes combined action has been shown to increase the effect (i.e. growth inhibition of yeast) observed in the bioassay (Koshinsky *et al*. 1992a,b). However, for exposure to mixtures of more than two mycotoxins as present in food products there is a clear lack of information and no test strategy has been brought forward. This paper proposes such a strategy. It is based on bioassays and it uses a tiered approach with a sequence of statistical experimental designs. These statistical designs have earlier been shown to be useful to detect possible interactions between food contaminants in laboratory animals which were orally exposed to mixtures of the compounds at several dose levels (Narotsky *et al*., 1995; Groten *et al*., 1996; Study 4.2).

In the present study, statistical designs will be employed (1) to assess the combined action of mixtures of five *Fusarium* mycotoxins and (2) to test the applicability of sequential

experimentation to identify interactions between individual compounds in the mixture. Due to the less restrictive nature of *in vitro* testing we were challenged to extend our statistical analysis from two dose levels as used during the *in vivo* Study 4.2 towards full dose-response curves. Three experimental stages were performed using a DNA synthesis-inhibition assay with mammalian cells. The objective was to investigate the combined action of five *Fusarium* mycotoxins. In the first stage, the mycotoxins were applied either singly or jointly in a constant ratio; it was meant to establish either departure from or agreement with additivity of the compounds at 4 different dose levels. This experiment forms the basis for the follow-up experiment, which was formulated to assess the sources of non-addition with a central composite design comprising 27 dose combinations. Finally, two interactions of particular interest were further studied to verify the extent and nature of the interactions in separate full factorial designs with two factors at a time.

Materials and methods

Preparation of mycotoxins

T-2 toxin (T-2), Deoxynivalenol (DON), Nivalenol (NIV), Zearalenone (ZEA), and Fumonisin B_1 (FB_1) were purchased from Sigma, St. Louis, USA. Stock solutions of 2 mg/ml of the mycotoxins in dimethyl sulfoxid (DMSO; Sigma, St. Louis, USA) were stored at -20 ^{o}C. On the day of exposure, various culture media, containing either a single mycotoxin or a mixture of all five mycotoxins, were prepared from the stock solutions. The final concentration of DMSO in the culture medium was 0.5%. In preliminary studies we have verified that this is a non-toxic concentration of the solvent in the present bioassay. Control medium was prepared with 0.5% DMSO without mycotoxins in the culture medium.

Cell lines and culture conditions

Mouse fibroblast L-929 cells were obtained from ATCC (CCL1 NCTC, clone 929 strain L). The cells were cultured in Dulbeco's MEM supplemented with heat-inactivated fetal calf serum (10% v/v) and 20 g/ml gentamicin (Gibco, Grand Island, USA) at 37 ^{o}C under a humidified atmosphere with 5% CO_2 in 75 cm^3 culture flasks. Cultures were refreshed three times a week.

DNA synthesis inhibition assay

Cells were seeded in 96-well tissue culture plates at a density of $5x10^3$ cells in 100 µl per well. After 24 h, the culture medium was removed by aspiration and replaced with 100 µl medium containing either a single mycotoxin or a mixture of mycotoxins, dissolved in DMSO. After an exposure period of 24 h, 10 µl of Methyl-H^3-thymidine (Amersham, Aylesbury, UK 1.85 MBq/ml) was added to each well in the presence of the mycotoxins and incubated for 5 h under the normal culture condition. Then the cells were trypsinised and harvested with an automated sample harvester (Harvester 96, Tomtec, Orange, USA) on glass fiber filters. The radioactivity was counted in a 1450 Microbeta-counter (Wallac, Turku, Finland). The percentage DNA synthesis was expressed as the ratio of radioactivity in exposed cells to the radioactivity in control cells. This value was subtracted from 100% to obtain the DNA synthesis inhibition.

Design of the study

The study was built up in three stages. First, an experiment with a mixture of all 5 mycotoxins at a constant ratio was performed to detect possible non-additivity for the mixture in its entirety. The mixture was studied at various concentrations. Secondly, the five mycotoxins were studied with a varying ratio of their concentrations in order to screen for interactive effects between the mycotoxins in the mixtures. This stage involved a fractional factorial design. Finally, two interactions of particular interest were studied more fully in two separate experiments using full factorial designs. All three stages included the assessment of the dose-response relationships of individual mycotoxins. More detailed information on each of the stages follows.

Stage 1. Detection. Five *Fusarium* mycotoxins were applied either individually or in combination at four concentrations each. The highest concentration of T2, NIV, ZEA, and DON was intended to result in an inhibition of DNA synthesis of approximately 30%. This concentration was established in a series of range-finding studies prior to stage 1.

The highest concentration of FB_1 was set at 4µg/ml since in previous experiments, FB_1 had no influence on DNA synthesis of L929 cells up to this test concentration. The highest concentration for DON, T2, ZEA and NIV were respectively 0.2 µg/ml, 2 ng/ml, 4 µg/ml, and 0.12 µg/ml. The remaining concentrations were chosen to be 1/8, 1/4, or 1/2, of the respective highest concentrations. Mixtures with the highest concentration of each of the mycotoxins were applied either undiluted, or at a dilution of 1/8, 1/4, or 1/2 of the initial concentration.

A total of four tissue culture plates were used. Each dilution was allotted to a separate plate. All dilutions were tested in 6 wells. A solution of 0.5% DMSO served as a negative control. The negative control was tested in 6 wells on every tissue culture plate.

Stage 2. Screening. In stage 2, mixtures were tested using five equi-distant concentrations of each of the mycotoxins. For T-2, ZEA, DON, and NIV, the highest concentration was chosen such that the effect induced by a single chemical would not exceed 30% DNA synthesis inhibition, as based on the results of stage 1. FB_1 is not cytotoxic to the cells up to the concentration of 10 mg/ml. The maximum dose of FB_1 was chosen such that DNA synthesis of L929 cells would increase.

The exposure levels of each mycotoxin were coded from -2 to +2, as shown in Table 41, in which +2 corresponds with the highest concentration tested (cf. the 30% effect level), and -2 stands for the lowest effect level (cf. the 5% effect level. The equi-distant dose levels were employed in a central composite design in order to examine possible 2 way-interactions in the mixtures of five *Fusarium* mycotoxins. A total of 27 combinations, given in Table 42, were chosen from 5^5 possible combinations of a full factorial design with 5 chemicals studied with five concentrations each. Sixteen combinations (cube points) were derived from a half-fraction of a 2^5 factorial design (Box *et al*., 1978, pp. 376-385); each mycotoxin was applied either at its -1 or at its +1 level. One combination (centre point) had all mycotoxins at their middle level. Ten combinations (star points) had four

Table 41. Dose levels of mycotoxins (µg/ml) used in the central composite design and the full factorial design

Coded	*Actual*				
	T-2	DON	NIV	ZEA	FB_1
-2	$0.5x10^{-3}$	0.05	0.040	1.9	1.0
-1	$1.1x10^{-3}$	0.11	0.054	2.4	1.75
0	$1.7x10^{-3}$	0.17	0.068	2.9	2.5
+1	$2.3x10^{-3}$	0.23	0.082	3.4	3.25
+2	$2.9x10^{-3}$	0.29	0.096	3.9	4.0

mycotoxins at their middle level, and the remaining one either at its highest of at its lowest level. The star points were included to provide information on curvilinear trend for each of the mycotoxins. A general reference for central composite designs is Box and Draper (1987; pp. 508-514).

A total of 8 plates were used. The design was replicated four times. Each replication involved two separate plates (there were eight plates in total). A plate was either used for star points or for cube and centre points. Each plate also had dose response wells for single mycotoxins, using all five concentrations of all five mycotoxins on every plate.

Stage 3. Confirmation. Two interactions of particular interest (NIV-T2 and FB_1-ZEA), as revealed by the central composite design were studied further in separate full factorial designs with two factors (two mycotoxins) using the same dose levels as in Stage 2. The three mycotoxins which were not varied in a particular design were applied at a their middle concentration. Each design was replicated four times, each replication involving a separate plate.

Table 42. DNA synthesis (% of control synthesis) in the central composite design[a]

	T-2	DON	NIV	ZEA	FB_1	fitted (addition)	[b]	observed	fitted (interaction)
16 cube points	-	-	-	-	+	57.47		54.46	55.61
	+	-	-	-	-	47.81		48.61	46.34
	-	+	-	-	-	49.55		48.46	46.61
	+	+	-	-	+	42.15	<	47.41	47.42
	-	-	+	-	-	52.58		51.73	55.43
	+	-	+	-	+	44.74	<	50.10	48.63
	-	+	+	-	+	46.36		50.47	50.52
	+	+	+	-	-	38.57		36.54	36.30
	-	-	-	+	-	47.74		51.03	50.58
	+	-	-	+	+	40.62		41.76	41.29
	-	+	-	+	+	42.09		41.05	41.54.
	+	+	-	+	-	35.02	<<<	44.78	43.13
	-	-	+	+	+	44.67		44.17	43.14
	+	-	+	+	-	37.16		37.58	38.63
	-	+	+	+	-	38.51		38.89	40.12
	+	+	+	+	+	32.77	>	27.42	28.25
10 star points	-2	0	0	0	0	48.69		52.04	50.29.
	+2	0	0	0	0	34.47		34.36	36.76
	0	-2	0	0	0	48.31		47.42	47.24.
	0	+2	0	0	0	36.73		35.08	36.08
	0	0	-2	0	0	45.53		48.37	51.60
	0	0	+2	0	0	38.97		43.13	41.98
	0	0	0	-2	0	51.94		53.82	54.18
	0	0	0	+2	0	38.25	<<	46.05	46.18
	0	0	0	0	-2	43.12		37.63	38.06
	0	0	0	0	+2	47.13		46.68	47.01
1 centre point	0	0	0	0	0	42.12	<	46.91	46.55

[a] Fitted values from additivity surface, observed values, and fitted values from interaction model for each mixture of 5 mycotoxins. See Table 41 for coding of dose levels. [b] Observed data significantly less than additive: < (P<0.05), << (P<0.01), <<< (P<0.001); more than additive : > (P<0.05).

Statistical Analysis

Generalised Linear Models. The radioactivity of the harvested cells was related to the dosage of the mycotoxins using a Generalised Linear Model (GLM) with a variance proportional to the mean and a logarithmic link function. (See McCullagh and Nelder, 1989, for a comprehensive reference.) The negative controls of the respective plates were included in the analysis as points with (uncoded) dose 0 for each of the mycotoxins. A specific model term accounted for plate differences.

As an example, the following equation might result from fitting observed radioactivity (cpm) to the dosage of DON alone in stage 1:

$$\ln(\text{cpm, DON plate i}) = \ln(\text{cpm, neg. control, plate i}) - 0.0775\ \text{DON} - 0.3052\ \text{DON}^2$$

Here i is an index for the plate. The logarithmic function links the mean value of the counts per minute with a function that is linear in its parameters. Parameter estimates are 0.0775 (parameter for DON), and -0.3052 (parameter for DON^2); 4 parameters, the estimates of which are not given, account for the different plates.

The proportionality of the variance of the observations with their means defines a Poisson-type distribution, which is often used for analyzing counted data. The Poisson distribution itself has a constant of proportionality of 1. Often, however, this constant is greater than one (over-dispersion; McCullagh and Nelder, 1989). We estimated the constant for each of the data sets to be analyzed.

The results of each GLM can be expressed as percentages of control activity by rewriting the model. For example, the above equation may be rewritten as

$$\ln(\%\ \text{of control}) = \ln(100) - 0.0775\ \text{DON} - 0.3052\ \text{DON}^2$$

This formula gives the fitted percentage activity. The observed percentage activity was either the ratio of the mean observed activity of a plate to the fitted control activity of a plate or the geometric average of this ratio across the plates.

Additivity Surface. For stages 1 and 2, a single additivity surface was fitted to all the data from the negative controls and the exposures to the individual mycotoxins. Initially, a model was fitted with linear, quadratic and cubic terms for each of the mycotoxins. Backward elimination of model terms (Draper and Smith, 1981, pp. 305-306) was applied to reduce the degree of the polynomial needed for each of the mycotoxins as much as possible. The procedure uses large-sample t tests appropriate to the GLM to assess whether the equation describes the data significantly worse when a term is dropped from the model.

Departures from additivity in the mixtures. The data of the mixtures from stages 1 and 2, respectively, were compared to predictions from the fitted additivity surface, as follows. A modified additivity surface was fitted to the joint data of mixtures and single exposures, by including so called indicator variables in the GLM. There is one indicator variable for each particular mix. These variables all have the value 0 for single exposures and 1 for the particular mix. The statistical significance of the discrepancy between observed and predicted values of the individual mixes under the additivity assumption was assessed by large-sample t tests of the indicator variables.

Response surfaces in the studies. In stages 2 and 3, the joint data from the mixtures and the controls were used to fit second-order equations to describe the relation between activity and the applied doses of each of the mycotoxins. These equations have linear and quadratic coefficients for the individual mycotoxins and product terms for the interaction between two mycotoxins. As an example, the following equation would be used for a study when two compounds, say A and B, are tested:

$$\ln(\%\ \text{of control}) = \ln(100) + d + a_1.x + a_2.x^2 + b_1.y + b_2.y^2 + c.x.y,$$

where x and y are the doses of compounds A and B, respectively, d is an unknown parameter accounting for a possible discrepancy between the mixes and the controls, a_1 and a_2 are unknown parameters associated with the main effect of compound A, b_1 and b_2 are unknown parameters associated with the main effect of compound B, and c is the parameter associated with interactive effect between compound A and B. If the product term between A and B is not significant, the model can be considered as response-additive. The parameters (coefficients) are estimated under the above GLM. It is also possible to test for lack of fit of these equations by calculating cubic coefficients of the mycotoxins. A particular cubic coefficient is, however, aliased with product terms of the dosage of the corresponding mycotoxin with the squared dosage of the other mycotoxins (Box and Draper, 1987; pp. 316-322). Therefore, if there is a 'cubic' lack of fit, there are several equations that describe the data equally well.

Results

Stage 1. Detection

The results of the first experiment are summarised in Fig. 11. When tested individually,

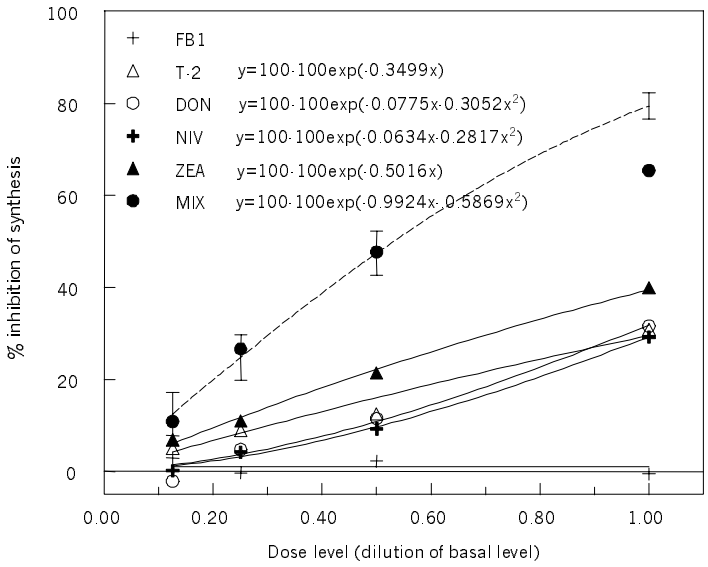


Fig. 11. Inhibition of DNA synthesis in cells treated with T-2 toxin (T-2), Deoxynivalenol (DON), Nivalenol (NIV), Zearalenone (ZEA) and Fumonisin B_1 (FB_1) applied singly or in a 5 compound mixture (MIX) in the Detection stage. Dose levels are expressed as dilutions of the following base-level concentrations. FB_1: 1 µg/ml; NIV: 0.12 µg/ml; DON: 0.2 µg/ml; ZEA: 4 µg/ml; T-2: 2 ng/ml. Marked points: observed inhibition; lines: predicted values under additivity. Error bars: 95% confidence intervals for discrepancies between observed and predicted values.

four of the mycotoxins showed a clear dose-dependent inhibition of the DNA synthesis. At the highest concentration, this was close to the intended level of 30%. FB_1 did not inhibit the DNA synthesis at the concentrations chosen. The cells exposed to a mixture of mycotoxins showed a more pronounced inhibition of the DNA synthesis than corresponding single exposures. The observed inhibitions of the DNA synthesis of the cells exposed to a mixture of mycotoxins for the dilutions 1/2, 1/4, and 1/8 were compatible with the effect predicted under response addition, the equation of which is given in Fig. 11. The mixture of mycotoxins tested at the highest dose, however, behaved statistically significant less than additive ($P < 0.001$).

Stage 2. Screening

For each of the single mycotoxins in the central composite design, the generalised linear model was converted into a percentage of the response and the final equations and fitted dose response curves are shown in Fig. 12. A dose related decrease in the DNA synthesis was observed in cells exposed to DON, NIV, T-2 and ZEA. FB_1 showed an increase in DNA synthesis at the highest concentration.

Cells exposed to a mixture of mycotoxins showed a DNA synthesis ranging between 54% and 27% of the controls. In Table 42, the observed data on DNA synthesis was compared with fitted values from an additivity surface using the equations given in Fig. 12. Five of the mixture combinations were significantly less than additive and one combination was more than additive. The observed data were also compared with fitted values from a response surface calculated from mixtures and negative controls alone (Table 42). Discrepancies between observed and fitted values were at most 3.8%. We consider this to

Table 43. Response surface results from the Central Composite Design

description	regression coefficient (s.e.)	statistical significance
Constant	-6.15 (0.014)	***
T-2	0.356 (0.132)	**
DON	4.06 (1.32)	**
NIV	23.7 (5.91)	***
ZEA	3.20 (1.36)	*
FB_1	1.464 (0.378)	***
$T\text{-}2^2$	-0.0551 (0.0262)	*
DON^2	-8.33 (2.65)	**
ZEA^2	-1.006 (0.472)	*
FB_1^2	-0.386 (0.145)	**
T-2-NIV	-4.41 (1.41)	**
DON-NIV	-34.5 (14.1)	*
NIV-ZEA	-4.84 (1.69)	**
ZEA-FB_1	-0.1467 (0.0317)	***
ZEA^3	0.1239 (0.0543)	*
FB_1^3	0.0458 (0.0187)	*

*** $P<0.001$; ** $P<0.01$; * $P<0.05$.

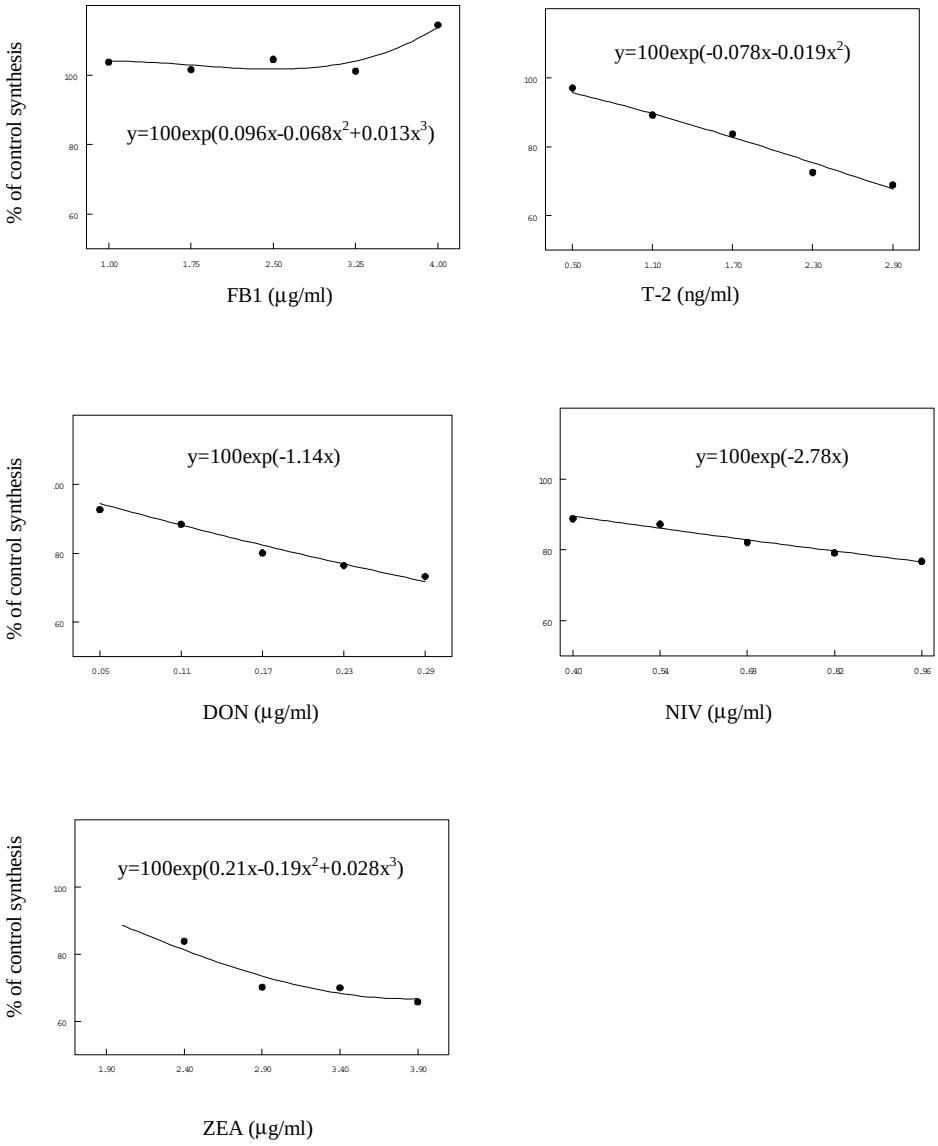


Fig. 12. DNA synthesis in cells exposed to a single mycotoxin in the Screening stage. Observed mean data (dots) and fitted dose response curves for FB_1, T-2, DON, NIV, and ZEA.

be satisfactory.

The response surface results are summarised in Table 43. This Table shows four significant interactions, viz. the interactions between ZEA and FB_1 ($p<0.001$), T-2 and NIV ($p<0.01$), DON and NIV ($p<0.05$), and NIV and ZEA ($p<0.01$), respectively. Drawing dose-response curves of ZEA at either the lowest or the highest level of FB1 as shown in Fig. 13a characterized the interaction between ZEA and FB1 more fully. There is no clear effect of ZEA on the DNA synthesis at the -2 level of FB_1. However at the +2 level of FB_1, ZEA decreases the DNA synthesis in a dose-dependent manner. Thus FB_1 affects the DNA synthesis in the presence of ZEA in a more than additive way. We conclude that the interaction between ZEA and FB_1 points in the direction of synergism.

Fig. 13b shows the interaction of NIV with T-2. The effect of the counter part of NIV, *viz.* DON, is not clearly observed at the -2 level of NIV, but becomes obvious at the +2 level of NIV. There, the counter part inhibits DNA synthesis in a dose dependent manner. Therefore the interactions with NIV can be considered as synergistic. The remaining two interactions (DON-NIV and NIV-ZEA), although not shown here explicitly, resulted in a similar pattern.

Stage 3. Confirmation

The synergistic interactions between ZEA and FB_1, and between NIV and T-2, as observed in the screening phase, were each separately studied in a full factorial design. Table 44 gives the observed DNA synthesis of the cells exposed to single mycotoxins, together with the corresponding effects as observed in stage 2. There was a dose-related decrease in DNA synthesis of cells exposed to NIV, ZEA, and T-2. FB_1 increased DNA synthesis at the +1 and +2 level. The results from stages 2 and 3 clearly indicate that the susceptibility of the cells towards the mycotoxins differed between the two subsequent experiments.

Mixtures of the ZEA / FB_1 did not show significant interaction (Table 45). The dose-response curves of joint action between these two mycotoxins are given in Fig. 14a. However, mixtures of T-2 and NIV as tested in a full factorial design did reveal a significant interaction. Dose-reposes curves are given in Fig. 14b and the equation of the response surface analysis is shown in Table 46. The effect of T-2 is not obvious at the lowest dose of NIV, but at the highest level, T-2 decreases DNA synthesis. Thus, the interaction between NIV and T2 on DNA synthesis inhibition in L929 cells can be considered as synergistic. The interaction profiles observed for T2 and NIV in the full factorial design were similar to the ones found at stage 2 in the fractional factorial design.

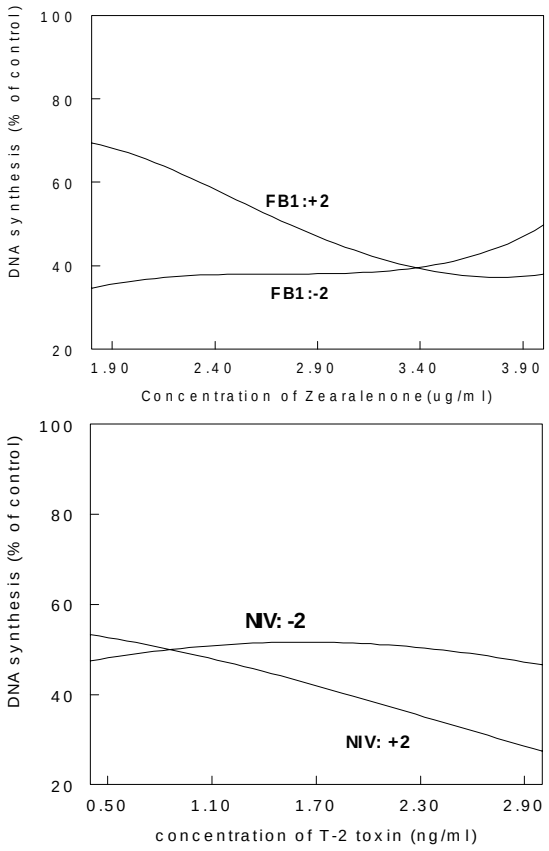


Fig. 13. Interaction between ZEA and FB_1, and between T-2 and NIV, on inhibition of DNA synthesis in cells exposed to mixtures of five mycotoxins in the Screening stage. See Table 43 for dose-response equation. For the three mycotoxins not visualised in a particular graph, the middle dose level was assumed.

Table 44. Susceptibilities of L929 cells to mycotoxins in screening and confirmation stages[a]

mycotoxin	dose level	central composite design	full factorial design
ZEA	-2	89.91	97.77
	-1	83.77	79.42
	0	70.13	81.32
	+1	69.96	79.90
	+2	65.71	81.40
FB1	-2	103.74	104.25
	-1	101.57	100.20
	0	104.83	101.92
	+1	101.14	110.31
	+2	114.44	109.08
T-2	-2	97.06	91.59
	-1	89.18	84.74
	0	83.67	85.12
	+1	72.51	82.64
	+2	68.90	82.56
NIV	-2	88.78	82.77
	-1	87.23	81.69
	0	82.06	76.22
	+1	79.11	69.28
	+2	76.72	71.23

[a]Susceptibility expressed as percent of control DNA synthesis.

Table 45. Response surface results from the 5x5 full factorial design (ZEA x FB_1)

description	regression coefficient (s.e.)	statistical significance
Constant	10.531 (0.0117)	***
ZEA	0.271 (0.105)	**
ZEA^2	-0.221 (0.0721)	**
ZEA^3	0.0354 (0.0119)	**
FB1	0.0212 (0.005)	***

*** $P<0.001$; ** $P<0.01$.

Discussion

The aim of the present study was to investigate the joint action of five *Fusarium* mycotoxins and to develop a test strategy to examine interactions among them. For this purpose, we used an *in vitro* assay measuring the inhibition of DNA synthesis in mammalian L929 cells in a tiered approach with several sequentially applied statistical designs to test joint actions or interaction of mixtures of mycotoxins at various concentrations.

In vitro cell culture assays have recently contributed to many aspects of mycotoxin research, including the quantitative determination, identification and characterisation of toxicity of mycotoxins. The present study shows that T-2 toxin, Deoxynivalenol, Nivalenol and Zearalenone dose-dependently inhibit the DNA synthesis of the cells in a 24 h incubation period. The cells were more sensitive towards type A Trichothecenes (T-2) than towards type B Trichothecenes (DON, NIV); Type B Trichothecenes were more cytotoxic than ZEA. This hierarchy in cytotoxicity of the *Fusarium* species under study seems to be in agreement with previous reports (Porcher *et al*., 1987; Visconti *et al*, 1991; Ueno *et al*., 1995). FB_1 behaves differently compared to the other mycotoxins tested and it increased DNA synthesis of L929 cells with about 10% at the highest concentration. As described for Swiss 3T3 cells, this stimulation of the DNA synthesis after FB_1 exposure might be associated with the accumulation of cellular sphingoid bases (Schroeder *et al*., 1994). Generally sphingolipids and their sphingoid bases are known to play a critical role in cell growth and differentiation. In fact, exogenously added sphingoid bases stimulate DNA synthesis in Swiss 3T3 cells. Similarly, FB_1 stimulates DNA synthesis in these cells, because it blocks *de novo* synthesis of sphingolipids. More in particular, it inhibits an enzyme that catalyses sphingolipid synthesis from sphingoid bases, which leads to accumulation of sphingoid bases in the cells.

Table 46. Response surface results from the 5x5 full factorial design (T-2 x NIV)

description	regression coefficient (s.e.)	statistical significance
constant	-0.2544 (0.0934)	***
T-2	-2.156 (0.536)	*
$T\text{-}2^2$	0.495 (0.119)	***
$T\text{-}2^3$	-0.1498 (0.0237)	***
T-2-NIV	-2.156 (0.543)	*
NIV	-0.047 (0.913)	*NS*

*** P<0.001; ** P<0.01;* P<0.05; *NS* P>0.05.

Our bioassay was applied to assess the toxicity of mixtures with a sequence of statistical experimental designs to assess joint action, to screen for specific interactions, and to confirm potential interactions. In the detection stage, mixtures of mycotoxins were tested at constant ratios and the effect of the mixture was compared to the effects of the individual compounds. This type of mixture study can be regarded as the simplest, and most frequently applied test design to investigate the presence of interaction in mixtures (Groten *et al.*, 1999). We initially assumed that the joint action of the mycotoxins could be described by response addition, which reflects the independent joint action of dissimilar

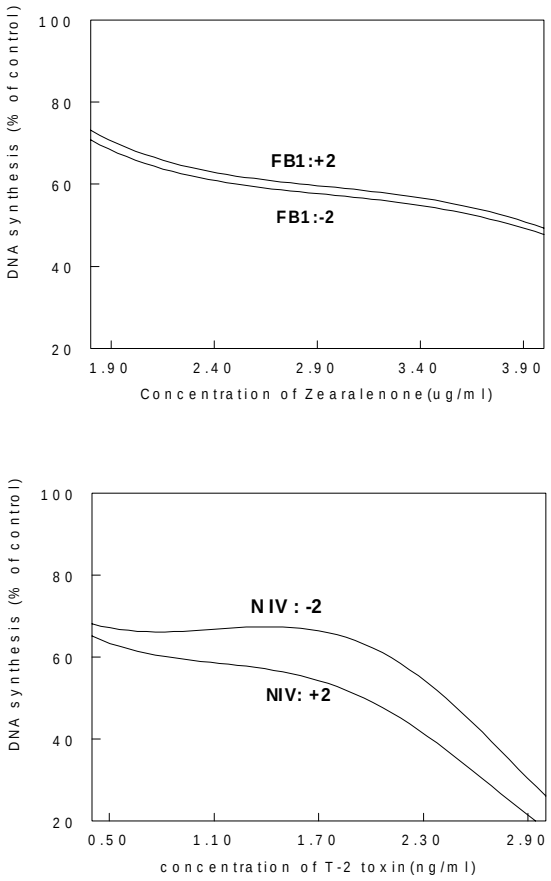


Fig. 14. Joint action of ZEA and FB_1, and between T-2 and NIV on inhibition of DNA synthesis in cells exposed to mixtures of five mycotoxins in separate 5 x 5 factorial designs. See Tables 45 and 46 for dose-response equations. The three mycotoxins not studied in a particular design were applied at their middle dose level.

acting compounds (Groten *et al*., 1999). This is because the mycotoxins used here belong to three classes with known distinct mechanisms of action. Even within the Trichothecenes class, which have similar chemical structures, the compounds may not always exert their action in a common way. T-2 and HT-2, for example, show different dose-response curves for growth inhibition in yeast cells (Koshinsky *et al*., 1992a).

The results from our first step, the detection stage, confirmed our additivity assumption for the three lowest dose levels. However, at the highest dose level the observed effect of the mixtures was less pronounced than predicted on the basis of response additivity. We do not have a sound explanation for this dose-dependent interaction, but it supports the view that mixture research should be carried out preferably at whole dose-response curves (Feron *et al*., 1995; Cassee *et al*., 1998; Groten, 2000).

The detection stage merely shows an overall absence of additivity in relatively toxic dose ranges; one is not able to identify specific interactions between pairs of mycotoxins.
Screening for interactive effects between chemicals in a defined chemical mixture requires statistical designs to optimise clarity and manageability of the toxicological experiments. In step 2, the screening stage, we used multi-factor experiments that were specifically designed for the exploration of these interactive effects and to establish response surfaces. This type of experimental design involves a lower number of experimental points than with a full multi-way factorial design. The central composite design used here is among the most frequently used response-surface designs (Study 4.1). These designs permit a screening for interactions in complex mixtures in (part of) the established response-surface area. Significant interactions between compounds in a certain dose-response area as found in the screening can be confirmed in a follow-up experiment by using full factorial designs. This stepwise approach was followed in the stages 2 and 3 of the present study.

A striking observation in the screening stage was the less-than-additivity of five of the mixes in combination with synergism between several pairs of mycotoxins. Note, however, that the synergism is apparent when comparing the observed syntheses of the mixes to each other, while the less-than-additivity is apparent when comparing the observed synthesis of an individual mix with a predicted synthesis from exposures to single mycotoxins. Thus, the interaction between the mycotoxins is more complex than we initially expected. At an overall level, they inhibit each other's action as shown by the less-than-additivity. But within the less than additive response, pairs of mycotoxins enhance each other's action. The first finding is important for the risk assessment of toxic compounds, while the second finding is important for the further clarification of the mechanism of the compounds at toxic doses.

Note that it is not possible to use our particular central composite design for interactions between more than two compounds. If this were suspected, we should use the full 2^5 factorial design as cube points, because this permits estimation of all interactions between three, four or five factors.

Traditionally, central composite designs, such as the one used in our stage 2, are a follow-up to fractionated two-level factorial designs. The latter designs are used to detect whether there are interactions at all. We used a slightly different approach. Our initial experiment already suggested that there were one or more active interactions. So we directly applied a central composite design. The results permit several alternative interpretations. In particular, the joint action of ZEA and FB_1 could not be identified well. A 'cubic' term in FB_1, for example, may be real, but it may also indicate a complex type of

two-factor interaction (Box and Draper, 1987, pp. 316-322). Thus, Fig. 13a and Table 43 present just one of several possible interpretations of the results from the screening stage. For this reason, a five by five full factorial design for ZEA and FB_1 was conducted subsequently.

The interaction between ZEA and FB_1 detected in the screening stage was not confirmed in Stage 3. This emphasises the importance of this confirmation stage, which fully concentrates on the interaction of particular interest. From the screening stage we knew that the joint action was not well defined. We were nevertheless surprised not to find some kind of interaction in the confirmation stage. Possibly, this is due to the difference in cell susceptibilities to ZEA and FB_1 between the experiments. Even though we have used the same batch of cells in both stages, it appears to be difficult to obtain constant sensitivities of the cells.

A further finding from the screening stage was the presence of three synergistic interactions with NIV. The interaction between T-2 and NIV was confirmed in the follow-up stage; the others were not tested further. To our knowledge, no study of the combined effect between T-2 and NIV has been reported before. A synergistic interaction between DON and NIV was reported previously *in vitro* using the yeast-bioassay (Madhyastha *et al*., 1994). Regarding other binary mixture of Trichothecenes assessed by *in vitro* systems, the joint effect appears to be synergistic in some cases and additive in others (Terse *et al* 1993; Madhyasth *et al*., 1994). The variation in combined effects of the Trichothecenes may be ascribed to differences in the ratios of Trichothecenes in the mixture, or to differences in effect levels (Koshinsky *et al*., 1992a; Jones *et al*., 1995). More mechanistic cellular physiological studies need to be carried out to investigate the extent and nature of these interactions, which may include the rate of transport of mycotoxins into cells, changes in affinity for the active binding site and metabolism of mycotoxins by cytosol enzymes (Cawood *et al.*, 1994)

Our study suggests that mixtures of *Fusarium* mycotoxins mostly act additively in terms of inhibition of DNA synthesis for L929 cells, although our statistical design revealed some specific two-factor interactions. We have illustrated this by using known mixtures of mycotoxins. Follow-up studies from our laboratory have been undertaken to measure the inhibition of DNA synthesis capacity of unknown agricultural samples contaminated with mycotoxins. For barley samples, for example, we have shown that this bioassay is able to detect defined mixtures of Trichothecenes which were spiked in a concentration below the detection limit of the GC/ECD analysis. The final goal is to incorporate this bioassay in a test strategy to screen combinations of Trichothecenes in contaminated samples of food and raw materials. The bioassay may alert for the presence of mixtures of toxins and is complementary to the chemical analysis.

Our study shows that a tiered approach with a sequence of test stages is a suitable strategy of investigation on mixtures of compounds. Stage 1 is a very compact way of assessing joint action (i.e. additivity or non-additivity). The follow-up with the central composite design is useful for screening of specific interactions in the mixture. A verification experiment with a full factorial design should be performed as a follow-up to confirm the interactions detected in the screening stage. We conclude that several classes of mycotoxins when present simultaneously in a mixture may show interaction. The effect of the mixture can not be predicted solely on the basis of the effect of the individual compounds.

Chapter 5 Summary of practical applications

All the studies in the previous three chapters featured practical examples of statistical design of experiments. In the present chapter, these examples will be evaluated in the light of the availability of the particular statistical techniques that were used to reach the conclusions.

Study 2.1, titled *Design and analysis of a fractional $4^1 3^1 2^5$ split-plot experiment*, featured an experiment from molecular genetics. Its purpose was to identify factors that inluence the yield of DNA amplified in a chemical reaction. Originally, there were two pieces of equipment to be tested in conjunction with the operating temperature of the equipment, and five ingredients of the chemical reaction under investigation. A third piece of equipment was brought in shortly after the preparation of the experimental plan. The treatments tested with the new device were copied from those of one of the original devices. The experiment identified for all the devices the factors that affect the yield of the chemical reaction.

Two particular statistical techniques permitted the detection of the influential factors. Firstly, in the part of the experiment with the original two devices, care was taken to carry out a total of four runs with each of the devices, with a total of four individual chemical reactions in each of these runs. The minimum number of runs that is required with the original devices equals four; these are just the combinations of two devices and two operating temperatures. However, it would then be impossible to separate the random variation between the runs of a device from systematic effects of device and operating temperature.

The second statistical technique to permit the detection of influential factors is the analysis of the results in pairs of devices. Indeed, there are just four runs carried out with each of the devices. Thus, the evaluation of these runs of the single devices would be problematic.

Study 2.2, titled *Designing fractional two-level experiments with nested error structures*, had two practical examples. The first one was a chemical experiment. Its purpose was to find from a list of 14 potentially influential factors those that really affect the yield of a catalyst synthesised on gauze. In the experiment, three influential factors were identified.

The chemical experiment had three sources of random variation: there were several working days involved, there were several jars to treat the gauzes in, and there were various gauzes to be treated. A full assessment of the magnitude of these sources of random variation in conjunction with the assessment of the 14 potentially influential factors was not feasible, as there was a maximum of 32 chemical syntheses. For this reason, the random variation due to the different jars was made to coincide with the variation due to the different gauzes. That is, each of the four gauzes treated on a particular day was treated in a different jar.

The use of different jars for all of the gauzes permitted the evaluation of the systematic effects from the experiment in a set of seven and a set of 24 effect. (See the General Introduction for the division of the effects in separate sets.) The effects of each of the sets are plotted to separate influential effects from those that are not influential.

Another experimental arrangement would be to use two jars per day. However, this would result in a set of seven, a set of eight and a set of 16 effects. As it is easier to separate influential effects from inactive effects in a plot with 24 effects than in a plot of 16 effects, the former arrangement is to be preferred.

The second example in Study 2.2 was concerned with a cheese-making experiment. Here there were predetermined numbers of milk supplies (8), numbers of curds productions per milk supply (4) and numbers of sets of identically treated cheeses per curds production (4). There were 11 factors potentially affecting the quality of the cheeses. Two of these operated on milk supplies, five operated on curds productions and four operated on sets of cheeses. The purpose of the experiment to identify the really influential factors was fully accomplished. However, it was not easy to construct the statistical design. This is because the standard designs of investigating 11 factors in 128 runs are not compatible with the requirement to investigate 7 factors in a total of 32 curds productions. For this reason the author developed an experimental strategy to investigate a given number of experimental factors with a predetermined set of experimental material. This strategy was successfully applied in the cheese-making example.

Study 3.1, titled *Three robust scale estimators to judge unreplicated experiments*, used a published example from microbiology. Here, there was a screening of 19 ingredients potentially enhancing the yield of *Aspergillus niger* in solid state fermentation. The 19 main effects of the ingredients were evaluated with a standard error calculated from the same set. This revealed that one of the ingredients may affect the yield. The estimated standard error was much larger than the standard error calculated from repeated observations with the same treatment. This is a discrepancy that may help to pinpoint possible insufficiencies in the design of the original study. If the proposed analysis had been carried out subsequent to the original study, it could have saved the pursuing of fruitless efforts to optimise factors that do not affect the fermentation.

Study 3.2, titled *Assessment of some critical factors in the freezing techniques for the cryopreservation of precision-cut rat liver slices*, was focused on characterising the difference between two techniques that preserve rat liver slices for future investigation. Slices were frozen according to various protocols. The slices were then kept stored and finally thawed after storage. After thawing, several measurements were made to quantify the viability of the cells. The study consisted of two sub-studies. The initial one focused on detection of overall differences between two techniques applied under conditions that are otherwise identical. The viability measurements for which the results of the respective techniques differed were further investigated in the second sub-study. Here, four factors were defined that characterise the differences in the techniques. There were also two experiments carried out separately. It was shown that residual viability of liver slices after cryopreservation and subsequent culturing is mainly determined by freezing rate and the cryopreservation medium. One particular combination of these factors was identified as the most promising approach to successful freezing of rat liver slices.

As to the statistical techniques used to reach conclusions, the first sub-study used pairwise *t* tests to judge the difference between the two methods. The second sub-study aimed at detecting the critical factors explaining the differences with a two-level design with 5 factors. The design was seemingly replicated because of the presence of pairs of

identically treated liver slices. However, as these slices are treated simultaneously, the relevant standard error for the effects is the one between the pairs of slices. As the treatments were carried out only once, the statistical evaluation of the effects used the procedure for unreplicated designs presented in Study 3.1. More conventional statistical techniques would have assumed the absence of interactions between three or more factors. As our analysis had one clearly active interaction of this type, the application of our specific statistical technique paid off well.

In Study 3.3, titled *Choosing appropriate two-level experiments to detect location and dispersion effects*, the two-level experiment from Study 3.2 was used as a practical example. Here, the variation among pairs of identically treated liver slices was analysed with various statistical tests. There is evidence that the cryopreservation medium affects the variability of the resulting activity of an enzyme. At present, the biological relevance of this finding is not clear. However, the author hopes that it triggers discussion on this subject. He also feels that the study of variability of a response to some treatment is important in risk assessment of toxic compounds.

Study 4.1, *Statistical designs in combination toxicology: a matter of choice*, was a discussion of the practical relevance of statistical designs from industrial research to toxicology of mixtures. This study demonstrated that there are often several designs that meet the objectives of a toxicological study. The choice between the alternative designs is a matter of balancing complexity of experimental conduct against the gain in information useful for the toxicologist.

Study 4.2, *Subacute toxicity of a combination of nine chemicals in rats: detecting interactive effects with a two level factorial design*, presented an experiment to detect interaction beween pairs of chemical compounds applied in a mixture to rats. In the experiment, each of the chemicals was either absent in the mixture or present at its 'minimum-observed-adverse-effect level' (MOAEL). With a total of 16 different mixtures, the interactions that were likely to be active could be identified. The statistical design that permitted the identification was a standard fractional factorial design. If these designs had not been available, the identification of interactions between the compounds studied would require more conventional approaches. For example, the 36 possible interactions between the nine compounds could also be identified with 36 separate experiments with either or both of two compounds administered at its MOAEL. In conjunction with a control group, this would require a total of 3 x 36 + 1 = 109 experimental groups. It is evident that practical research here would not have been possible without the availability of fractional designs.

Study 4.3, *Statistically designed experiments in a tiered approach to screen mixtures of* Fusarium *mycotoxins for possible interactions*, proposed a sequence of experimental designs as a strategy to detect interactions. First, five *Fusarium* mycotoxins were applied in a mixture to cells grown in 12-wells plates. The response of interest was the rate at which the cells were synthesising new DNA. The mixture was studied at four dilutions. There were also exposures to single mycotoxins in concentrations also attained in the mixes. The author developed a statistical test to compare the activity of cells exposed to the mixes with

the activity predicted on the basis of the exposures to single mycotoxins. This test revealed that exposure to the highest concentration of the mix results in a cell-activity not compatible with addition of the effects of mycotoxins that are applied singly.

All the mixes in the first stage had the same ratio of the concentrations of the mycotoxins. In the second stage of the strategy, cells were exposed to mixes with various ratios. Each mycotoxin was applied at one of five concentrations. From the 3125 possible combinations, 27 were actually tried. The 27 specific combinations correspond with a central composite design. The experiment identified four interactions. However, statistical tests specific for the central composite design showed that these interactions could also be interpreted as complex curvilinearity in the main effects of the compounds. It is therefore essential to perform a follow-up. This was actually done for two of the possible interactions. One of the interactions was confirmed, and the other interaction was not confirmed.

The practical strategy for the detection of interactions in mixtures heavily relies on the statistical test for the detection of non-addition (Stage 1) and efficient experimental designs for screening of interactions (Stage 2). Thus, this study, like the preceding one, is a clear example of applied scientific research that would not have been possible without the existence of specific statistical methods.

References

Addelman, S. (1964) Some two-level factorial plans with split-plot confounding. *Technometrics* **6**: 253-258.

Aitkin, M. (1987) Modelling variance heterogeneity in normal regression using GLIM. *Applied Statistics* **36**: 332-339.

Appelman, L.M., Woutersen, R.A., Zwart, A., Falke, H.E., and Feron V.J. (1988) One-year inhalation toxicity study of formaldehyde in male rats with damaged or undamaged nasal mucosa. *Journal of Applied Toxicology* **8**: 85-90.

Bach, P.H., Vickers, A.E.M., Fisher, R., Baumann, A., Brittebo, E., Carlile, D.J., Koster, H.J., Lake, B.G., Salmon, F., Sawyer, T.W., and Skabinsky, G. (1996) The use of tissue slices in pharmacotoxicology studies. The report and recommendations of ECVAM workshop 20. *Alternatives to Laboratory Animals* **24**: 693-923.

Bartlett, M.S., and Kendall, D.G. (1946) The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Journal of the Royal Statistical Society B* **8**: 128-138.

Berenbaum, M.C. (1985) The expected effect of a combination of agents: the general solution. *Journal of Theoretical Biology* **114**:413-431.

Berger, W.W.A., van der Stap, J.G.M.M., and Kientz, C.E. (1985) Trichothecene production in liquid stationary cultures of *Fusarium tricinctum* NRRL 3299: Comparison of quantitative brine shrimp assay with physicochemical analysis. *Applied Environmental Microbiology* **50**: 656-662.

Bergman, B., and Hynén, A. (1997) Dispersion effects from unreplicated designs in the 2^{k-p} series. *Technometrics* **39**: 191-198.

Bhavanishanker, T.N., Ramesh, H.P., and Shantha, T. (1988) Dermal toxicity of Fusarium toxins in combinations. *Archives of Toxicology* **61**: 241-244.

Bingham, D.R., and Sitter, R.R. (1999) Minimum-aberration fractional split-plot designs. *Technometrics* **41**: 62-70.

Bisgaard, S. (1994a) A note on the definition of resolution for blocked 2^{k-p} designs. *Technometrics* **36**: 308-311.

Bisgaard, S. (1994b) Blocking generators for small 2^{k-p} designs. *Journal of Quality Technology* **26**: 288-296.

Bliss, C.I. (1939) The toxicity of poisons applied jointly. *Annals of Applied Biology* **26**: 585-615.

Box G.E.P., and Behnken, D.W. (1960) Some new three level designs for the study of quantitative variables. *Technometrics* **2**:455-475.

Box, G.E.P., and Draper, N.R. (1987) *Empirical model-building and response surfaces* (New York, Wiley).

Box G.E.P., and Hunter, J.S. (1957) Multifactor experimental designs for exploring response surfaces. *Annals of Mathematical Statistics* **28**:195-242.

Box, G.E.P., and Hunter, J.S. (1961) The 2^{k-p} fractional factorial designs I. *Technometrics* **3**: 311-352.

Box, G.E.P., and Jones, S. (1993) Split-plot designs for robust product experimentation. *Journal of Applied Statistics* **19**: 3-26.

Box, G.E.P., and Meyer, R.D. (1986) Dispersion effects from fractional designs. *Technometrics* **28**: 19-27.

Box G.E.P., and Wilson, K.B. (1951) On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society Ser. B* **13**:1-45

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

Buckle, A.E., and Sanders, M.F. (1990) An appraisal of bioassay methods for the detection of mycotoxins-a review. *Letters in Applied Microbiology* **10**: 155-160.

Carter, W.H., Jr., and Gennings, C. (1994) Analysis of chemical combinations: relating isobolograms to response surfaces. In *Toxicology of chemical mixtures* (R.S.H. Yang, editor; Academic Press, San Diego).

Cassee, F.R., Groten, J.P., Schoen, E.D., and Feron, V.J. (1996a) Sensory irritation to mixtures of formaldehyde, acrolein and acetaldehyde in rats. *Archives of Toxicology* **70**: 329-337

Cassee, F.R., Groten, J.P., Feron, V.J. (1996b) Changes in the nasal epithelium of rats exposed by inhalation to mixtures of formaldehyde, acetaldehyde and acrolein. *Fundamental and Applied Toxicology* **29**: 208- 218.

Cassee, F.R., Groten, J.P., van Bladeren, P.J., and Feron, V.J. (1998) Toxicological evaluation and risk assessment of chemical mixtures. *Critical Reviews in Toxicology* **28**: 73-101.

Cawood, M.E., Gelderblom, W.C.A., Alberts, J.F., and Snyman, S.D. (1994) Interaction of ^{14}C-labelled fumonisin B mycotoxins with primary rat hepatocyte cultures. *Food Chemical Toxicology* **32**: 627-632.

Chaturved, A.K. (1993) Toxicological evaluation of mixtures of ten widely used pesticides. *Journal of Applied Toxicology* **13**: 183-188.

Connor, W.S., and Young, S. (1961) *Fractional factorial designs for experiments with factors at two and three levels* (U.S. Dept. of Commerce, National Bureau of Standards, Applied Mathematics Series 58).

Coulombe Jr., R.A. (1993) Symposium: Biological action of mycotoxins. *Journal of Dairy Science* **76**: 880-891.

Cox, D.R. (1958) *Planning of experiments* (New York, Wiley).

Cox, D.R. (1970) *The analysis of binary data* (Chapman and Hall, London).

Daniel, C. (1959) Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics* **1**: 311-341.

Daniel, C. (1976) *Applications of statistics to industrial experimentation* (New York: Wiley).

Day, S.H., Nicoll-Griffith, D.A., and Silva, J.M. (1999) Cryopreservation of rat and human liver slices by rapid freezing. *Cryobiology* **38**: 154-159.

Diaz, G. J., Squires, E. J., Julian, R. J., and Boermans, H. J. (1994) Individual and combined effects of T-2 toxin and DAS in laying hens. *British Poultry Science* **35**: 393-405.

Dirven, H.A.A.M, van den Broek, P.H.H., and Jongeneelen, F.J. (1990) Effect of di(ethyl hexyl)phthalate on enzyme activity levels in liver and serum of rats. *Toxicology* **65**: 199-207.

Doko, M. B., Canet, C., Brown, N., Sydenham, E. W., Mpuchane S., and Siame, B. A. (1996) Natural co-occurrence of Fumonisins and Zearalenone in cereal and cereal-based

foods from Eastern and Southern Africa. *Journal of Agricultural and Food Chemistry* **44**: 3240-3243.

Dong, F. (1993) On the identification of active contrasts in unreplicated fractional factorials. *Statistica Sinica* **3**: 209-217.

Draper, N.R., and Smith, H. (1981). *Applied regression analysis* (2nd edition; Wiley and Sons, New York).

Draper, N.R., and Smith, H. (1998). *Applied Regression Analysis* (3d edition, Wiley, New York).

Draper, N.R., Davis, T.P., Pozueta, L., and Grove, D.M. (1994) Isolation of degrees of freedom for Box-Behnken designs. *Technometrics* **36**: 283-291.

Dunnett, C.W. (1964) New tables for multiple comparisons with a control. *Biometrics* **20**: 482-491.

Engel, J., and Huele, A.F. (1996) A generalized linear modelling approach to robust design. *Technometrics* **38**: 365-373.

Ekins, S. (1996) Vitrification of precision-cut rat liver slices. *Cryo-letters* **17**: 7-14.

Ekins, S., Williams, J.A., Murray, G.I., Burke, M.D., Marchant, N.C., Engeset, J., and Hawksworth, G.M. (1996) Xenobiotic metabolism in rat, dog, and human precision-cut liver slices, freshly isolated hepatocytes, and vitrified precision-cut liver slices. *Drug Metabolism and Disposition* **24**: 990- 995.

Eriksson, L., Jonsson, J., Hellberg, S., Lindgren F., Skagerberg B., Sjostrom F., and Wold, S. (1991). A strategy for ranking environmentally occurring chemicals III. Multivariate quantitative structure-activity relationships for halogenated aliphatics. *Environmental and Toxicological Chemistry* **9**: 1339-1351

Fahy, G.M. (1990) Vitrification as an approach to organ preservation: Past, present, and future. In: *Cryopreservation and low temperature biology in blood transfusion* (C.Th. Smit Sibenga, P.C. Das, and H.T. Meryman, editors; Kluwer Academic Press, Amsterdam).

Feron, V.J., de Groot, A.P., Spanjers, M.T., and Til, H.P. (1973) An evaluation of the criterion "organ weight" under conditions of growth retardation. *Food and Chemical Toxicology* **11**: 85-94.

Feron, V.J., Woutersen, R.A., Arts, J.H.E., Cassee, F.R., de Vrijer, Fl., and van Bladeren P.J. (1992) Indoor air, a variable complex mixture: strategy for selection of (combinations of) chemicals with high health hazard potential. *Environmental Technology* **13**: 341-350.

Feron, V.J., Groten, J.P., Jonker D., Cassee, F.R., and van Bladeren, P.J. (1995a) Risk Assessment of simple (defined) mixtures of chemicals. In *Toxicology and Air Pollution: Risk Assessment* (V. Burgat-Sacaze, Descotes J., Gaussens P., and Leslie G.B., editors; Faculté de Médicine et de Pharmacie, Université de Bourgogne).

Feron, V.J., Groten, J.P., Jonker, D., Cassee F.R., and van Bladeren, P. J. (1995b) Toxicology of chemical mixtures: challenges for today and the future. *Toxicology* **105**: 415-427.

Feron, V.J., Woutersen, R.A., Arts, J.H.E., Cassee, F.R., de Vrijer, Fl., and van Bladeren, P.J. (1995c) Safety evaluation of the mixture of chemicals at a specific workplace: theoretical considerations and a suggested two-step procedure. *Toxicology Letters* **76**: 47-55.

Fisher, R.L., Hasal, S.J., Sanuik, J.T., Scott, K.S., Gandolfi, A.J., and Brendel, K. (1993) Cold- and cryopreservation of human liver and kidney slices. *Cryobiology* **30**: 250-261.

Fisher, R.L., Putnam, C.W., Koep, L.J., Sipes, I.G., Gandolfi, A.J. and Brendel, K. (1991) Cryopreservation of rat and human liver slices. *Cryobiology* **28**: 131-142.

Fries, A., and Hunter, W.G. (1980) Minimum aberration 2^{k-p} designs. *Technometrics* **22**: 601-608.

Fuller, H.T., and Bisgaard, S. (1995) A comparison of dispersion effect identification methods from unreplicated two-level factorials (Report 132, University of Wisconsin-Madison, Center for Quality and Productivity Improvement).

Gennings, C., (1995) An efficient experimental design for detecting departure from additivity in mixtures of many chemicals. *Toxicology* **105**: 189-197.

Gessner, P. (1988) A straightforward method for the study of drug interactions: an isobolographic analysis primer. *Journal of the American College of Toxicology* **7**: 987-1012.

Glöckner, R., Steinmetzer, P., Drobner, C., and Müller, D. (1998) Application of cryopreserved precision-cut liver slices in pharmaco-toxicology - principles, literature data and own investigations with special reference to CYP1A1-mRNA induction. *Experimental and Toxicologic Pathology* **50**: 440-449.

Groten, J.P.(2000) Mixtures and Interactions. *Food and Chemical Toxicology* **38**: S65-S71.

Groten, J.P., Sinkeldam, E.J., Muys, T., Luten, J.B., and van Bladeren, P.J. (1991) Interaction of dietary Ca, P, Mg, Mn, Cu, Fe, Zn ans Se with the accumulation and oral toxicity of cadmium in rats. *Food and Chemical Toxicology* **29**: 249-258.

Groten, J.P., Schoen, E.D., and Feron, V.J. (1996) Use of factorial designs in combination toxicity studies. *Food and Chemical Toxicology* **34**: 1083-1089.

Groten, J.P., Cassee, F.R., van Bladeren, P.J., De Rosa, C., and Feron, V.J. (1999) Mixtures. In: *Toxicology* (H. Marquardt., S.G. Schafer, R, McClellan, and F. Welsch, editors; Academic Press, San Diego)

Haaland, P.D., and O'Connell, M.A. (1995) Inference for contrast-saturated fractional factorials. *Technometrics* **37**: 82-93.

Heindel, J.J., Chapin, R.E., Gulati, D.K., George, J.D., Price, C.J., Marr, M.C., Myers, C.B., Barnes, L.H., Fail, P.A., Grizzle, T.B., Schwetz, B.A., and Yang, R.S.H. (1994) Assessment of the reproductive and developmental toxicity of pesticide/fertilizer mixtures based on confirmed pesticide contamination in California and Iowa groundwater. *Fundamental and Applied Toxicology* **22**: 605-621.

Hissin, P.J., and Hilf, R.A. (1976) Fluorimetric method for determination of oxidized and reduced glutathione in tissues. *Analytical Biochemistry* **74**: 214-226.

Hochberg Y., and Tamhane, A.C. (1987) *Multiple comparison procedures* (John Wiley and Sons, New York)

Huang, P., Chen, D., and Voelkel, J.O. (1998): Minimum-aberration two-level split-plot designs. *Technometrics* **40**: 314-326.

Huele, A.F., Schoen, E.D., and Steeman, R.A. (submitted) A note on REML estimation in joint modelling of mean and dispersion.

Iersel, M.L.P.S. van, Ploemen, J-P.H.T.M., Struik, I., van Amersfoort, C., Keyzer, A.E., Schefferlie, J.G., and van Bladeren, P.J. (1996) Inhibition of glutathione S-transferase activity in human melanoma cells by α,β-unsaturated carbonyl derivates. Effects of

acrolein, cinnamaldehyde, citral, crotonaldehyde, curcumin, ethacrynic acid, and *trans*-2-hexenal. *Chemico-Biological Interactions* **102**: 117-132.

Ito, N., Hasegawa, K., Imaida K., Kurat Y., Hagiwara, A., and Shirai, T. (1995) Effect of ingestion of 20 pesticides in combination at acceptable daily intake levels on rat liver carcinogenesis *Food and Chemical Toxicology* **33**: 159-163.

Jones, T. J., Koshinsky, H. A., and Khachatourians, G. G. (1995) Effect of T-2 toxin and Verrucarin A in combination on *Kluyveromyces marxianus*. *Natural Toxins* **3**: 104-108.

Jonker, D., Woutersen, R.A., van Bladeren, P.J., Til, H.P., and Feron, V.J. (1990) 4-week oral toxicity study of a combination of eight chemicals in rats: comparison with the toxicity of the individual compounds. *Food and Chemical Toxicology* **28**: 623-631.

Jonker, D., Woutersen, R.A., van Bladeren, P.J., Til, H.P., and Feron, V.J. (1993) Subacute (4-wk) oral toxicity of a combination of four nephrotoxins in rats: comparison with the toxicity of the individual compounds. *Food and Chemical Toxicology* **31**: 125-136.

Kanter, R. de, and Koster, H.J. (1995) Cryopreservation of rat and monkey liver slices. *Alternatives to Laboratory Animals* **23**: 653-665.

Kanter, R. de, Olinga, P., Hof, I., de Jager, M., Verwillegen, W.A., Slooff, M.J., Koster, H.J., Meijer, D.K., and Groothuis, G.M. (1998) A rapid and simple method for cryopreservation of human liver slices. *Xenobiotica* **28**: 225-234.

Klooster, G.A.E. van 't, Blaauboer, B.J., Noordhoek, J., and van Miert, A.S.J.P.A.M. (1993) Cytochrome P450 induction and metabolism of alkoxyresorufins, ethylmorphine and testosterone in cultured hepatocytes from goats, sheep and cattle. *Biochemical Pharmacology* **46**: 1781-1790.

Koshinsky, H. A., and Khachatourians, G. G. (1992a) Trichothecene synergism, additivity, and antagonism: The significance of the maximally quiescent ratio. *Natural Toxins* **1**: 38- 47.

Koshinsky, H. A., and Khachatourians G. G. (1992b) Bioassay for Deoxynivalenol based on the interaction of T-2 toxin with Trichothecene mycotoxins. *Bulletin of Environmental Contamination and Toxicology* **49**: 246-251.

Krishnan, J., and Brodeur, J. (1991) Toxicological consequences of combined exposure to enviro nemental pollutants. *Archives of Complex Environmental Studies* **3**.

Krumdieck, C.L., Dos Santos, J.E., and Ho, K.J. (1980) A new instrument for the rapid preparation of tissue slices. *Analytical Biochemistry* **104**: 118-123.

Kubena, L.F., Edrington, T.S., Kamps-Holtzapple, C., Harvey, R.B., Elissalde, M.H., and Rottinghaus, G.E. (1995) Influence of Fumonisin B1, present in *Fusarium monoliforme* culture material, and T-2 toxin, on turkey poults. *Poultry Science* **74**, 306-313.

Lee, Y., and Nelder, J.A. (1998) Generalized Linear Models for the analysis of quality improvement experiments. *The Canadian Journal of Statistics* **26**: 95-105.

Lenth, R.V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics* **31**: 469-473.

Levin R.L., and Miller, T.W. (1981) An optimum method for the introduction or removal of permeable cryoprotectants: isolated cells. *Cryobiology* **18**: 32-48.

Littlefield, N.A., Farmer, J.H., Gaylor, D.W., and Sheldon, W.G. (1980a) Effects of dose, and time in a long-term low-dose carcinogenic study. *Journal of Environmental and Pathological Toxicology* **3**:17-34.

Littlefield, N.A., Greenman, D.L., and Farmer, J.H. (1980b) Effects of continuous, and discontinued exposure to 2-AAF on urinary bladder hyperplasia, and neoplasia. *Journal of Environmental and Pathological Toxicology* **3**:35-54.

Loh, W.-Y. (1992) Identification of active contrasts in unreplicated factorial experiments. *Computational Statistics and Data Analysis* **14**: 135-148.

Lorenzen, T.J., and Anderson,V.L. (1993) *Design of experiments: a no-name approach*. (Marcel Dekker, New York).

Maas, W.J.M., Leeman, W.R., Groten, J.P., and van de Sandt, J.J.M. (submitted) Cryopreservation of primary hepatocytes and precision-cut rat liver slices using a computer-controlled freezer.

Madhyasth, M.S., Marquardt, R.R., and Abramson D. (1994) Structure-activity relationships and interactions among Trichothecene mycotoxins as assessed by yeast bioassay. *Toxicon* **32**: 1147-1152.

Mazur, P. (1984) Freezing of living cells: mechanisms and implications. *American Journal of Physiology* **247**: 125-142.

McCullagh, P., and Nelder, J. A. (1989) *Generalized linear models*. (2nd edition, Chapman and Hall, London).

McLean, R.A. , and Anderson, V.L. (1984) *Applied factorial and fractional designs* (New York, Marcel Dekker).

Miller, A. (1997) Strip-plot configurations of fractional factorials. *Technometrics* **39**: 153-161.

Monod, H., and Bailey, R.A. (1992) Pseudofactors: normal use to improve design and facilitate analysis. *Applied Statistics* **41**: 317-336.

Montgomery, D.C. (1994) *Design and analysis of experiments* (3d edition, New York, Wiley).

Montgomery, D.C. (1997) *Design and analysis of experiments* (4th edition, New York, Wiley).

Mumtaz, M.M., DeRosa, C.T., and Durkin, P.R. (1994) Approaches and challenges in risk assessments of chemical mixtures. In: *Toxicology of Chemical Mixtures* (R.S.H. Yang, editor; Academic Press, San Diego).

Mumtaz, M.M., Sipes, G., and Yang, R.S.H. (1993) Risk assessment of chemical mixtures: biologic and toxicologic issues. *Fundamental and Applied Toxicology* **21**: 258-269.

Nair, V.N., and Pregibon, D. (1988) Analyzing dispersion effects from replicated factorial experiments. *Technometrics* **30**: 247-257.

Narotsky, M.G., Weller, E.A., Chinchilli, V.M., and Kevlock, R.J. (1995) Non-additive developmental toxicity in mixtures of trichloroethylene, (di(2-ethylhexyl)phthalate, and heptachlor in a 5x5x5 design. *Fundamental and Applied Toxicology* **27**: 203-216.

National Toxicology Program (1993) Toxicity studies of pesticides/fertilizer mixtures administered in drinking water to F344/N rats and B6C3F1 Mice. (NTP Technical Report Series No. 36, NTP Publication 93-3385).

Nelder, J.A., and Lee, Y. (1991) Generalized linear models for the analysis of Taguchi-type experiments. *Applied Stochastic Models and Data Analysis* **7**: 107-120.

Nelder, J.A., and Lee, Y. (1998) Joint modeling of mean and dispersion. In: Letters to the editor, with responses from B. Bergman and A. Hynén and from A.F. Huele and J. Engel. *Technometrics* **40**: 168-175.

Nesnow, S. (1994) Mechanistic linkage between DNA adducts, mutations in oncogenes, and tumorigenesis of carcinogenic polycyclic aromatic hydrocarbons in strain A/J mice. In: *Chemical mixtures and quantitative Risk Assessmen*t (Abstracts of The Second Annual HERL Symposium, November 7-10, 1994, Raleigh NC).

Norred, W.P., Bacon, C.W., Porter, J.K., and Voss, K.A. (1990) Inhibition of protein synthesis in rat primary hepatocytes by extracts of *Fusarium moniliforme*-contaminated corn. *Food and Chemical Toxicology* **28**: 89-94.

Oishi S., Oishi H., and Hiraga, K. (1979) The effect of food restriction for 4 weeks on common toxicity parameters in male rats. *Toxicological and Applied Pharmacology* **47**: 15-22.

Olinga, P., Meijer, D.K.F., Slooff, M.H.J., and Groothuis, G.M.M. (1997) Liver slices in *in vitro* pharmacotoxicology with special reference to the use of human liver tissue. *Toxicology in Vitro* **12**: 77-100.

Pan, G. (1999) The impact of unidentified location effects on dispersion-effects identification from unreplicated factorial designs. *Technometrics* **41**: 313-326.

Pazhayannur, P.V., and Bischof, J.C. (1997) Measurement and simulation of water transport during freezing in mammalian liver tissue. *Journal of Biomechanical Engineering* **119**: 269-277.

Plackett R.L., and Burman, J.P. (1946) The design of optimum multifactorial experiments, *Biometrika* **33**: 305-339.

Plackett, R.L., and Hewlett, P.S. (1952) Quantal responses to mixtures of poisons. *Journal of the Royal Statistical Society Ser. B* **14**: 141-163.

Porcher, J., Lafarge-Frayssinet, C., Frayssinet, C., Nurie, A., Melcion D., and Richard-Molard, D. (1987) Determination of cytotoxic Trichothecenes in corn by cell culture toxicity assay. *Journal of the Association of Official Analytical Chemists* **70**: 844-849.

Reubsaet, F.A.G., Veerkamp, J.H., Bukkens, S.G.F., Trijbels, J.M.F., and Monnens, L.A.H. (1988) Acyl-CoA oxidase activity and peroxisomal fatty acid oxidation in rat tissues. *Biochimica et Biophysica Acta* **958**: 434-442.

Rotter, B.A., Rotter, R.G., Thompson, B.K., and Trenholm, H.L. (1992) Investigations in the use of mice exposed to mycotoxins as a model for growing pigs. *Journal of Toxicology and Enviromental Health* **37**: 329-339.

Schoen, E.D. (1997) Cheese making with Genstat: a case study in design of industrial experiments. *Genstat Newsletter* **33**: 20-29.

Schroeder, J.J., Crane, H.M., Xia, J., Liotta D.C., and Merrill Jr., A.H. (1994) Disruption of sphingolipid metabolism and stimulation of DNA synthesis by Fumonisin B_1. *The Journal of Biological Chemistry* **269**: 3475-3481.

Scossa-Romano, D.A., Bickel, R.E., Zweifel, U., Reinhardt, C.A., Lüthy J.W., and Schlatter C.L. (1987) Fast and sensitive method for detection of Trichothecenes in maize by using protein synthesis inhibition in cultured fibroblasts. *Journal of the Association of Official Analytical Chemists* **70**: 129-132.

Seed J., Brown, R.P., Olin, S.S., and Foran, J.A. (1995) Chemical mixtures: Current risk assessment methodologies and future directions. *Registration Toxicology and Pharmacology* **22**: 76-94.

Singh, Y., Cooke, J.B., Hinton, D.E., and Miller, M.G. (1996) Trout liver slices for metabolism and toxicity studies. *Drug Metabolism and Disposition* **24**:7-14.

Smith, D.J., Pham, L.D., and Bischof, J.C. (1998) The effect of dimethylsulfoxide on the water transport response of rat liver tissue during freezing. *Cryo-letters* **19**: 343-354.

Southard, J.H., and Belzer, F.O. (1993) The University of Wisconsin organ preservation solution: components, comparisons, and modifications. *Transplantation Reviews* **7**: 176-190.

Srinivas, M.R.S., Chand, N., and Lonsane, B.K. (1994). Use of Plackett-Burman design for rapid screening of several nitrogen sources, growth/product promoters, minerals and enzyme inducers for the production of alpha-galactosidase by Aspergillus niger MRSS 234 in solid state fermentation system. *Bioprocess Engineering* **10**: 139-144.

Steel, G.G., and Peckham, M.J. (1979) Exploitable mechanisms in combined radiotherapy-chemotherapy: the concept of additivity. *International Journal of Radiation Oncology and Biological Physics* **5**: 85-91.

Sun, D.X., Wu, C.F.J., & Chen, Y. (1997) Optimal blocking schemes for 2^n and 2^{n-p} designs. *Technometrics* **39**: 298-307.

Svensgaard, D.J., and Hertzberg, R.C. (1994) Statistical methods for the toxicological evaluation of the additivity assumption as used in the environmental protection agency chemical mixture risk assessment guideline. In: *Toxicology of chemicals mixtures* (R.S.H. Yang, editor; Academic Press, San Diego).

Tanaka, T., Hasegawa, A., Yamnoto, S., Lee, U.S., Sugiura, Y., and Ueno, Y. (1988) Worlwide contamination of cereals by the *Fusarium* mycotoxins nivalenol, deoxynivalenol and zearalenone I. A survey of 19 countries. *Journal of Agricultural and Food Chemistry* **36**: 979-983.

Terse, P.S., Madhyastha, M.S., Zurovac, O., Stringfellow, D., Marquardt, R.R., and Kemppainen, W. (1993) Comparison of *in vitro* and *in vivo* biological activity of mycotoxins. *Toxicon* **31**: 913-919.

U.S. Environmental Protection Agency (1986). Guidelines for the health risk assessment of chemical mixtures. *Fed. Regist*. 51, 34014-34025.

Ueno, Y., Umemori, K., Niimi, E., Tanuma, S., Nagata, S., Sugamata, M., Ihara, T., Sekijima, M., Kawai, K., Ueno, I., and Tashiro, F. (1995) Induction of apoptosis by T-2 toxin and other natural toxins in HL-60 human promyelotic leukemia cells. *Natural Toxins* **3**: 129-137.

Verbyla, A. (1993) Modelling variance heterogeneity: residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society Ser. B* **55**: 493-508.

Visconti, A., Minervini, F., Lucivero, G., and Gambatesa, V. (1991) Cytotoxic and immunotoxic effect of Fusarium mycotoxins using a rapid colorimetric bioassay. *Mycopathologia* **113**: 181-186.

Wang, P.C. (1989) Tests for dispersion effects from orthogonal arrays. *Computational Statistics and Data Analysis* **8**: 109-117.

WHO (1987). *Air quality guidelines for Europe*. (WHO regional publications, European series No.23).

Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A. , and Tingey, S.V. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research* **18**: 6531-6535.

Wilmer, J.W.G.M., Woutersen, R.A., Appelman, L.M., Leeman W.R., and Feron,V.J. (1987) Subacute (4-week) inhalation toxicity study of formaldehyde in male rats: 8-hour

intermittent versus 8-hour continuous exposures. *Journal of Applied Toxicology* **7**: 15-16.

Wishnies, S.M., Parrish, A.R., Sipes, I.G., Gandolfi, A.J., Putnam, C.W., Krumdieck, C.L., and Brendel, K. (1991) Biotransformation activity in vitrified human liver slices. *Cryobiology* **28**: 216-226.

Wolff, K., Schoen, E.D., and Peters-Van Rijn, J. (1993) Optimizing the generation of random amplified polymorphic DNAs in chrysanthemum. *Theoretical and Applied Genetics* **86**: 1033-1037.

Woutersen, R.A., Appelman, L.M., Wilmer, J.W.G.M., Falke, H.E., and Feron, V.J. (1987) Sub chronic (13-week) inhalation toxicity study of formaldehyde in rats. *Journal of Applied Toxicology* **7**: 43-49.

Wu, C.F.J. (1989) Construction of 2^m4^n designs via a grouping scheme. *Annals of Statistics* **17**: 1880-1885.

Wu, C.F.J., and Zhang, R. (1993) Minimum aberration designs with two-level and four-level factors. *Biometrika* **80**: 203-209.

Yang, R.S.H. (1994a) Toxicology of chemical mixtures derived from hazardous waste sites or application of pesticides and fertilizers. In: *Toxicology of chemical mixtures* (R.S.H. Yang, editor; Academic Press, San Diego).

Yang, R.S.H. (1994b) Introduction to the toxicology of chemical mixtures. In: *Toxicology of Chemical Mixtures* (R.S.H. Yang, editor; Academic Press, San Diego).

Yang, R.S.H., and Rauckman, E.J. (1987) Toxicological studies of chemical mixtures of environmental concern at the National Toxicology Program: health effects of groundwater contaminants. *Toxicology* **47**: 15-34.

Yang, R.S.H., Hong, H.L., and Boorman, G.A. (1989). Toxicology of chemical mixtures: experimental approaches, underlying concepts, and some results. *Toxicology Letters* **49**: 183-197.

Appendix 1 Samenvatting

Bij het opstellen van het statistisch ontwerp van een experiment hebben we altijd te maken met randvoorwaarden die voortkomen uit vakgebied waarop het experiment betrekking heeft. Deze randvoorwaarden kunnen de ontwikkeling van nieuwe statistische ontwerp- en analysemethoden stimuleren. Anderzijds kan de beschikbaarheid van specifieke statistische ontwerpen experimenteel onderzoek stimuleren dat zonder die ontwerpen moeilijk uitvoerbaar zou zijn. De eis om aan randvoorwaarden te voldoen en de stimulans die van reeds ontwikkelde statistische ontwerpen kan uitgaan illustreren dat er vaak een interactie zal zijn tussen de statistische discipline van het ontwerpen van experimenten en het vakgebied van waaruit het experiment wordt opgezet. Het onderzoek dat in dit proefschrift is beschreven weerspiegelt de genoemde interactie. Niet alleen wordt er nieuwe statistische methodologie ontwikkeld om aan bepaalde randvoorwaarden de te voldoen, maar ook wordt ingegaan op de toepassing van bestaande statistische methodologie voor industriële experimenten op het vakgebied van de (mengsel)toxicologie. De nieuwe methodologie geeft oplossingen voor de omgang met meerdere bronnen van toevalsvariatie in een experiment, de inbedding van factoren op drie en op vier niveaus in een proefontwerp met factoren op twee niveaus en de analyse van effecten in ongerepliceerde experimenten met formele procedures. De bestaande methodologie stelt toxicologen in staat om het proefdiergebruik, de duur van een experiment en de kosten van experimenteren terug te dringen.

Hoofdstuk 2 behandelt uitbreidingen van statistische ontwerpmethoden. Er worden twee studies beschreven. Studie 2.1 betreft een onderzoek naar ontwerp en analyse van fractionele experimenten met één factor op vier niveaus, één factor op drie niveaus en p factoren op twee niveaus. In een fractioneel experiment is er een onderlinge verwevenheid van effecten. Zo kan men in een bepaald soort fractioneel experiment bij voorbeeld de som van twee interacties te weten te komen, maar niet beide interacties afzonderlijk. Het probleem bij Studie 2.1 is om een geschikt ontwerp te maken met de factoren op vier, op drie en op twee niveaus, en de onderlinge verwevenheid van de effecten met een geschikt criterium te beoordelen. In aansluiting daarop moet ook een geëigende analyse gevonden worden.

Gewoonlijk proberen we voor een ontwerp met een beperkt aantal van de factoren op meer dan twee niveaus een aansluiting te maken op bestaande ontwerpen met uitsluitend factoren op twee niveaus. Voor de bestaande ontwerpen zijn namelijk al beoordeling- en analysemethoden gepubliceerd. Voor de factor op vier niveaus kunnen we de genoemde aansluiting maken door twee factoren op twee niveaus te combineren. De factor op drie niveaus wordt veelal geconstrueerd door een tussenfactor op vier niveaus aan te maken en twee van die niveaus vervolgens samen te nemen. In Studie 2.1 wordt echter een afwijkend constructieprincipe gehanteerd. Uit een ontwerp met uitsluitend factoren op twee niveaus wordt een factor gekozen die twee van de drie niveaus van de toekomstige factor belichaamt. Vervolgens wordt het ontwerp aangevuld met een halve fractie van datzelfde ontwerp, waarbij die halve fractie wordt uitgevoerd onder het nog ontbrekende niveau van de factor op drie niveaus.

Voor de beoordeling van het overall ontwerp stellen we voor om een opsplitsing te maken naar de drie mogelijke paren van niveaus van de factor op drie niveaus. Vervolgens kan de verwevenheid van de effecten in de drie deelontwerpen van het overall ontwerp worden beoordeeld. We laten zien dat met de voorgestelde procedure onderscheid gemaakt kan worden in betere en slechtere overall ontwerpen.

Voor de analyse stellen we voor om bestaande analysemethoden voor factoren op twee niveaus toe te passen op elk van de hierboven genoemde deelontwerpen. Hierbij wordt de factor op vier niveaus ontbonden in twee 'pseudofactoren' van elk twee niveaus (Daniel, 1959).

Een aspect van Studie 2.1 dat nog niet is genoemd betreft de aanpassing van een ontwerp voor een fractioneel experiment met een enkele bron van toevalsvariatie aan een praktijksituatie waarbij twee bronnen van toevalsvariatie een rol spelen. Die bronnen zijn bovendien genest: bij elke realisatie van de eerste toevalsvariabele vinden steeds weer nieuwe realisaties van de tweede toevalsvariabele plaats.

In Studie 2.2 werken we het werken met geneste bronnen van toevalsvariatie verder uit door een algemene strategie voor te stellen om factoren op twee niveaus te bestuderen in aanwezigheid van geneste bronnen van toevalsvariatie. We bekijken twee deelproblemen. Ten eerste behandelen we maatregelen in het experimentele protocol om ervoor te zorgen dat het aantal afzonderlijk te onderscheiden bronnen van toevalsvariatie beperkt blijft.

Het tweede deelprobleem dat in Studie 2.2 aan de orde komt is het vinden van een set behandelingen bij een gegeven schema voor de bronnen van toevalsvariatie. Ter illustratie gebruiken we een experiment uit de kaasbereiding. Hierbij worden uit elk van 32 bakken wrongel vier partijen kaas geproduceerd. Voor het ontwerp maken we gebruik van twee deelontwerpen die we aan elkaar koppelen. Het voorbeeld betreft een deelontwerp voor zeven factoren die tussen de bakken wrongel worden gevarieerd en een deelontwerp voor vier factoren die tussen de partijen kaas binnen de wrongelbakken worden gevarieerd. Elk van de deelontwerpen komt uit een bestaande catalogus van experimenten met factoren op twee niveaus.

De resultaten van de experimenten uit Hoofdstuk 2 zijn beoordeeld met grafische weergaven van de effecten. De beoordeling heeft daarmee een duidelijk subjectieve component. Meer formele beoordelingsprocedures komen in Hoofdstuk 3 aan de orde. Deze procedures zijn bestemd voor ongerepliceerde experimenten. De fractionele experimenten uit Hoofdstuk 2 zijn daar een deelverzameling van.

Hoofdstuk 3 bestaat uit drie studies. In Studie 3.1 worden effecten uit ongerepliceerde experimenten getoetst worden met een standaardfout, geconstrueerd vanuit de effecten zelf. De standaardfout mag niet verontreinigd worden door actieve effecten, want dan wordt hij te groot. We zeggen dan dat de standaardfout robuust moet zijn tegen contaminatie met actieve effecten. Studie 3.1 bespreekt drie schatters van de standaardfout die elk op een iets verschillende manier robuust zijn tegen contaminatie. In de studie stellen we als richtlijn voor om de keuze van de schatter af te laten hangen van de mate van verwevenheid van de effecten uit het gehanteerde proefschema. We veronderstellen daarbij dat een geringe mate van verwevenheid gebruikt wordt bij experimenten waarbij men veel actieve effecten verwacht en een hogere mate van verwevenheid voor experimenten waarbij men slechts enkele actieve effecten verwacht. Volgens de richtlijn moet de schatter voor de

standaardfout van effecten meer robuust zijn bij het eerstgenoemde dan bij het laatstgenoemde type experimenten.

De schatters uit Studie 3.1 behoeven constanten die afhangen van het aantal effecten in een experiment. Ongerepliceerde experimenten zoals die uit Hoofdstuk 2 hebben daarbij een complicatie: de effecten zijn op te splitsen volgens het aantal bronnen van toevalsvariatie dat invloed heeft op de effecten. Dit resulteert in twee of meer sets. Eén ervan bevat effecten die slechts gevoelig zijn voor één bron van toevalsvariatie, terwijl de andere sets gevoelig zin voor twee of meer bronnen. De standaardfout van de effecten zal dus per set verschillend zijn. Daarom moet men per set afzonderlijk de standaardfout bepalen. Dat zou eenvoudig zijn als de bijbehorende constanten makkelijk beschikbaar zijn. Studie 3.1 verhoogt de praktische bruikbaarheid van de drie besproken schatters door de benodigde constanten te geven voor effect sets van zeven tot en met 127 effecten. De constanten zijn verkregen met computersimulatie.

De resultaten van Studie 3.1 zijn in dit proefschrift verder gebruikt om de 31 effecten te analyseren van een ontwerp met vijf factoren op twee niveaus (Studie 3.2). De effecten komen uit een experiment waarin wordt nagegaan welke van de verschillen tussen twee invriesprocedures voor weefselplakjes bepalend is voor het verschil in de levensvatbaarheid van de cellen na ontdooiing. De invriessnelheid en het gebruikte invriesmedium voor de plakjes bleken het resultaat te bepalen. Eén specifieke combinatie van deze factoren komt het meest in aanmerking als basis voor de ontwikkeling van verdere verfijningen.

De studie over invriesprocedures laat zien dat de detectieprocedure van Studie 3.1 op zichzelf al waarde heeft. De procedure kan echter ook gebruikt worden als de eerste stap van een methodiek om na te gaan of factoren invloed hebben op de grootte van de toevalsvariatie. Factoren waarvoor dit inderdaad het geval is kunnen gebruikt worden om spreiding van industriële producten terug te brengen of om een meer homogeen resultaat van experimentele procedures te krijgen. De detectieprocedure voor effecten op toevalsvariatie, of dispersie-effecten, wordt uitgewerkt in Studie 3.3, waarbij het reeds genoemde experiment uit Studie 3.2 als voorbeeld dient.

De procedure die in Studie 3.3 wordt voorgesteld heeft als eerste stap de detectie van locatie-effecten met een robuuste schatter voor de standaardfout van deze effecten. De detectie van dispersie-effecten vindt in de tweede stap van de procedure plaats. In de Studie wordt aangetoond dat de methoden voor de tweede stap onbruikbaar zijn als hoofdeffecten zijn verweven met twee-factor interacties. Ook zijn ze onbruikbaar als er interacties actief zijn. Met computersimulaties van de detectieprocedure in een breed scala van experimenten wordt aangetoond dat de voorwaarden waaronder een redelijke detectie van dispersie-effecten plaatsvindt in ongerepliceerde experimenten te restrictief zijn voor een praktisch bruikbare procedure. In gerepliceerde experimenten kan een enkel groot dispersie-effect gedetecteerd worden met 32 waarnemingen en twee grote effecten met 64 waarnemingen.

In Hoofdstuk 4 wordt ingegaan op de mogelijkheden die statistische ontwerpen van experimenten uit de industrie bieden aan onderzoek naar de toxiciteit van mengsels. In dit hoofdstuk worden drie studies beschreven. Studie 4.1 brengt naar voren wat de merites van de voorgestelde ontwerpen van experimenten zijn en wijst erop dat bij een gegeven toxicologische vraagstelling meerdere ontwerpen mogelijk zijn. Welk ontwerp uiteindelijk gekozen wordt is een zaak van het afwegen van praktische uitvoerbaarheid en de winst aan

nuttige informatie die uit een ingewikkelder alternatief kan worden gehaald. De keuze moet in een samenspraak tussen statisticus en toxicoloog worden gemaakt.

Eén van de voorbeelden uit Studie 4.1 wordt gebruikt om het nut te illustreren van fractionele ontwerpen ter detectie van interacties tussen componenten van een mengsel. Het voorbeeld wordt in Studie 4.2 verder uitgewerkt. Het betreft een experiment met 16 mengsels waarmee het effect van negen componenten bestudeerd wordt. In de gebruikte mengsels waren de componenten of afwezig of aanwezig op het zogenoemde Minimum Observed Adverse Effect Level . Het ontwerp laat de bepaling toe van negen hoofdeffecten en zes combinaties van twee-factor interacties. Van de 29 voornaamste responsvariabelen lieten er 16 activiteit van één of meer van deze combinaties zien. Welke interactie uit zo'n combinatie naar alle waarschijnlijkheid actief is kon worden opgemaakt uit informatie over hoofdeffecten en resultaten van eerder onderzoek.

Het fractioneel experiment van Studie 4.2 werd uitgevoerd zonder dat er experimentele aanwijzingen van interacties bekend waren. Een aantrekkelijk alternatief zou kunnen zijn om eerst een experiment uit te voeren om aan te tonen dat interacties een rol spelen, en te vervolgen met een experiment om na te gaan welke van de mogelijke interacties actief zijn. Deze notie vormt de kern van de strategie die in studie 4.3 wordt voorgesteld. De strategie wordt geïllustreerd met een reeks van experimenten om interacties tussen vijf *Fusarium* mycotoxinen te identificeren. In het eerste experiment worden cellen blootgesteld aan een reeks verdunningen van een mengsel van de vijf mycotoxinen. Ook blootstellingen aan individuele mycotoxinen zijn in het experiment opgenomen. Er wordt nagegaan of de schade die door het mengsel wordt veroorzaakt te verklaren valt uit de schades die de individuele mycotoxinen veroorzaken. Dit blijkt niet het geval te zijn. Hiermee is aangetoond dat de mycotoxinen interactie vertonen. In een vervolgexperiment wordt een zogenoemd centraal samengesteld ontwerp toegepast. Hieruit worden lineaire en kwadratische effecten en interacties van de mycotoxinen op de celschade geïdentificeerd. Met behulp van aanvullende statistische toetsen is vervolgens nagegaan of er niet nog meer modeltermen nodig zijn om de gegevens goed te beschrijven. Uit die toetsen wordt duidelijk dat er inderdaad extra termen nodig zijn, maar niet welke dat precies zijn. Ter opheldering worden volledige 5 x 5 factoriële experimenten voorgesteld voor elk van de gevonden interacties. Twee ervan zijn daadwerkelijk uitgevoerd. Deze bevestigen één van de interacties, maar de andere interactie wordt niet teruggevonden. De Studie toont aan dat effecten van de mengsels niet voorspeld kunnen worden uit die van de individuele componenten.

In Hoofdstuk 5, tenslotte, wordt besproken wat de specifieke statistische technieken hebben bijgedragen aan de resultaten van de praktijkvoorbeelden uit de diverse studies.

Appendix 2 Dankwoord

Ik heb de weg die naar dit proefschrift leidt niet alleen gelopen. Van de vele mensen die me vergezelden wil ik hier graag enkele noemen.

Ik werd op weg geholpen doordat mijn ouders trots op me waren. Ze zijn dat trouwens nog steeds. Hun warmte betekent veel voor mij.

Tijdens mijn biologiestudie aan de VU ben ik voor de statistiek gewonnen door Hans Jager met een memorabel college over de statistische toets. Hans liet me ook zien dat α en β in één persoon kunnen samengaan. Dat was voor mij een geruststellende gedachte. Ten behoeve van de toepassing van statistiek in de microbiologie zorgde Henk van Verseveld voor een werkbaar klimaat, ook waar ik wel eens te veel op standaard tijden de schaarse apparatuur bezette.

Wim van der Steen houdt me al jaren gezelschap op de weg die naar dit proefschrift leidt, maar ook op wegen naar andere bestemmingen. Toen ik nog op de VU studeerde schaakten we met elkaar, dronken voortreffelijke (rode) wijn met elkaar, en schreef ik mijn eerste schrijfsels, waarop hij verwoestende kritiek uitbracht. Zodoende hield hij me gaande. Later bestonden de gezamenlijke evenementen uit bridgen en wandelen. De vele gesprekken hielden 'mijn wetenschappelijk geweten wakker' en relativeerden het belang van de wetenschap van de meetbare werkelijkheid. Dat voelt lente-achtig aan. Wim, ik ben blij dat ik je ken. Het laatste glaasje Sambuca is voorlopig nog niet gedronken.

Dik Habbema van het Instituut Maatschappelijke Gezondheidszorg aan de Erasmus Universiteit te Rotterdam bewerkstelligde met zijn vraag of ik het allemaal nog wel leuk vond een terugkeer naar toegepaste statistiek. Ik ben hem daar nog steeds erkentelijk voor. Paul van der Maas van hetzelfde instituut zei me bij mijn afscheid dat hij het zou betreuren als ik geheel voor de wetenschap verloren zou gaan. Ik ben hem dankbaar voor deze woorden; ik ben ervan overtuigd dat ze van invloed zijn geweest op mijn verdere ontwikkeling.

Op de afdeling Statistiek van ITI-TNO, later de afdeling Toegepaste Statistiek van TNO TPD, zorgden Monique Bakker, Marion van den Bol, Peter Defize, Giljam Derksen, Pieter Marres en Jan Telman voor die typische bèta dakje humor waar een statisticus bij kan gedijen. Met instemming van deze personen werd ik in staat gesteld mijn wetenschappelijke slagkracht te vergroten door als stagebegeleider op te treden van Enrico Kaul, Peter van den Berg en Volkert Siersma van de TU Eindhoven. Ik ben Enrico, Peter en Volkert erkentelijk voor de stimulans die er van onze samenwerking uitging.

Met Kirsten Wolff, destijds van TNO Voeding, schreef ik in een sprankelende openheid Wolff *et al*. (1993) en Schoen and Wolff (1997). Kirsten, bedankt voor de samenwerking.

John Groten van TNO Voeding te Zeist is een groot voorstander van het gebruik van *designed* experiments in de mengseltoxicologie. Ik denk met veel plezier terug aan de mondelinge discussies die in de diverse studies van hoofdstuk 4 zijn uitgekristalliseerd. Zijn betrokkenheid bij mijn promotie door stukken van mijn proefschrift van commentaar te voorzien waardeer ik zeer.

Study 4.3 was conducted with Osamu Tajima of Kirin Brewery, Ltd., Japan as the first author. I admire his ability to work with the tiny wells of a 96-wells plate. His accuracy continued to astonish me. I also admire his persistence to obtain a pretty accurate

appreciation of statistical design and analysis; this subject was not well known to him at the start of our co-operation. Osamu, I hope that our efforts will be rewarded by the publication of our joint article, and I thank you for your kind permission to use the material of the mycotoxin study for my PhD thesis.

Van Wilfred Maas en Inge de Graaf, respectievelijk van TNO Voeding te Zeist en Solvay te Weesp, heb ik ondervonden hoe een kameraadschappelijke samenwerking eruit kan zien. Onze cryo-in-de-kroeg bijeenkomsten waren wetenschappelijk en culinair een belevenis. Dit smaakt naar meer, Wilfred en Inge.

Paul van der Laan, mijn eerste promotor, overtuigde me ervan dat ik best vanuit de statistiek kon promoveren met een accent op toepassingen in de toxicologie. Dat vond ik fijn, want daardoor kon ik zowel mijn statistische als mijn toxicologische manuscripten in de strijd werpen, en daarmee in het magische jaar 2000 promoveren. Vic Feron was geheel onbaatzuchtig bereid om van een potentiële eerste promotor naar een zekere tweede promotor te veranderen. Vic, ik dank je voor je enthousiasme en de wijsheid waarmee je me in het promotietraject hebt begeleid.

Jan Dijkstra, mijn copromotor, recenseerde in 1998 het stapeltje van mijn manuscripten en artikelen om te beoordelen of ‘er een promotie in zat’. Zijn oordeel ‘met tempo promoveren’ bemoedigde me om er vaart achter te zetten. Zijn eigenzinnig taalgebruik en de ruimte die er bij hem is voor vakgebieden buiten mathematische statistiek vind ik verfrissend. Ik denk ook dat Toegepaste Statistiek daar wel bij vaart.

Emiel van Berkum is als expert op het gebied van design of experiments al geruime tijd betrokken bij mijn promotietraject. Ik waardeer zijn betrokkenheid zeer, en ik hoop dat we nog eens tijd zullen vinden om gezamenlijke publicaties te schrijven.

Willem Seinen en Peter Sander dank ik voor hun bereidheid om in de kerncommissie zitting te nemen en zich te verdiepen in het manuscript van het proefschrift.

Bas Engel heeft me in een tijd dat ik het spoor bijster dreigde te raken weer op de goede weg teruggezet. Ik weet niet of ik heel veel met je concrete raadgevingen heb gedaan, Bas, maar, parbleu, je luisterend oor deed al wonderen.

En dan is het nu de tijd om mijn gezin te bedanken. Lieve Pepijn en Judith, het laatste jaar belemmerde ik het gebruik van computerspelletjes doordat ik vaak op zolder aan het promoveren was. Gelukkig neemt Mirjam de komende jaren die rol over door op zolder de boeken in te duiken voor haar studie Theologie. Dan kan ik mooi met jullie gaan schaken en afwassen.

Lieve Mirjam, zonder jouw steun en organisatietalent was het mij niet mogelijk geweest zo veel tijd aan het schrijven van artikelen te wijden. Ik heb iemand eens bij een trouwerij horen zeggen dat je je echtgenoot hoog moet houden. Dat heb je gedaan, want ik was vaak op de tweede verdieping te vinden. Ik hoop de komende jaren hetzelfde voor jou te doen. Dat we elkaar hoog houden geeft me diepe voldoening.

Appendix 3 Curriculum vitae

Eric Dick Schoen werd op 30 augustus 1956 in Zaandam geboren. In 1974 behaalde hij het diploma Gymnasium B aan de Christelijke Scholengemeenschap Pascal te Amsterdam. In hetzelfde jaar begon hij met de studie biologie aan de Vrije Universiteit, eveneens te Amsterdam. Hij behaalde in het najaar van 1978 het kandidaatsdiploma biologie en scheikunde met wiskunde en natuurkunde (cum laude). In de doctoraalfase van de biologiestudie volgde Eric als hoofdvak theoretische biologie en als bijvakken microbiologie en moleculaire genetica. Verder was hij betrokken bij het geven van onderwijs in statistiek, algemene methodologie en toegepaste wiskunde aan voorkandidaatsstudenten biologie. Ook behaalde hij de eerstegraads onderwijsbevoegdheid in de biologie. In januari 1983 sloot hij de studie af met het behalen van het doctoraaldiploma (cum laude).

In 1983 en 1984 was Eric in dienst bij de Vrije Universiteit als deeltijd wetenschappelijk medewerker. In die periode zette hij zijn onderwijsactiviteiten voort en volgde hij een methodologische training bij Wim van der Steen. Dit resulteerde in zijn eerste Engelstalige publicatie, met als onderwerp de toepassing van statistische methodologie in de microbiologie.

Van 1985 tot halverwege 1987 was Eric wetenschappelijk medewerker bij het Instituut Maatschappelijke Gezondheidszorg aan de Erasmus Universiteit te Rotterdam. Daar had hij als taak het bouwen van een rekenmodel om de effecten na te gaan die de systematische opsporing van hoge bloeddruk zou kunnen hebben op de verwachte levensduur van de Nederlandse bevolking.

In juni 1987 trad Eric als statistisch consultant in dienst bij TNO. De afdeling Statistiek waar hij te werk werd gesteld was toen onderdeel van het Instituut voor Toegepaste Informatica TNO. Later werd de afdeling toegevoegd aan TNO TPD.

Bij TNO raakte Eric gefascineerd door fractionele ontwerpen van experimenten. Om zijn inzichten uit te dragen formuleerde hij een cursus Design of Experiments voor onderzoekers. Hij heeft de cursus ruim twintig maal gegeven en voorzag daarmee mede in een persoonlijke behoefte aan het geven van onderwijs.

De artikelen waar dit proefschrift op gebaseerd is ontstonden uit de drang om praktijktoepassingen van systematisch experimenteren in een algemener kader te plaatsen en te publiceren. Het onderzoeksdeel van de activiteiten van Eric bij TNO werd vanaf 1996 periodiek versterkt door afstudeerstudenten c.q. een postdoctorale stagiaire van de TU Eindhoven. Eric trad daarbij op als begeleider vanuit TNO, Jan Dijkstra was de begeleider vanuit de TU Eindhoven. Een deel van de ontwikkelde software van de afstudeerstudenten is momenteel opgenomen in de procedurebibliotheek van het statistische pakket Genstat.

TNO stelde Eric in staat om internationale contacten aan te knopen. Met name genoemd moeten worden een werkbezoek aan professor John Nelder van het Imperial College in Londen (najaar 1996) en een werkbezoek aan Derek Bingham van de Universiteit van Michigan te Ann Arbor en aan Randy Sitter die daar toen te gast was (najaar 1999). Momenteel is Eric met de laatstgenoemde twee personen bezig een catalogus van fractionele split-plot experimenten met minimum aberratie te maken.

Stellingen bij Schoen (2000)

Schoen, E.D. (2000) *Issues in applying statistical design of experiments, with special reference to toxicology of mixtures*. Proefschrift Technische Universiteit Eindhoven.

1. Het proefdiergebruik voor de beoordeling van de acute inhalatietoxiciteit van een nieuw chemisch product kan worden teruggebracht door de LC_{50} studies voor de diverse voorgeschreven blootstellingsduren niet op zichzelf te laten staan, maar onder te brengen in één studie om de relatie tussen blootstellingsconcentratie, blootstellingsduur en sterfte te bepalen.

   Zwart, A., Arts, J.H.E., Klokman-Houweling, J.M., and Schoen, E.D. (1990) Determination of concentration-time-mortality relationships to replace LC50 values. *Inhalation Toxicology* **2**: 105-117.

2. Als men in een microbiologisch onderzoek wil aantonen dat de groei van een organisme door een nutriënt beperkt wordt bij toediening in een bepaalde hoeveelheid, dan moet men in één en hetzelfde experiment ook de groei bij hogere en of lagere hoeveelheden van het nutriënt onderzoeken onder overigens gelijkblijvende omstandigheden.

   Schoen, E.D., Jager, J.C., van Verseveld, H.W., and Stouthamer, A.H. (1985) Statistical analysis of growth limitations in *Paracoccus denitrificans*: an experiment with a completely randomized two-way factorial design with replications. *Antonie van Leeuwenhoek* **51**: 11-24.

3. De volgende 2^{k-p} proefopzetten hebben voor een hoofdeffectenmodel $\frac{1}{2}N-1$ residuele vrijheidsgraden, met $N = 2^{k-p}$: alle 2^{k-p} proefopzetten die geconstrueerd kunnen worden door aan een $2^{(k-1)-p}$ opzet die met hoofdeffecten verzadigd is zijn spiegelbeeld toe te voegen en een extra factor in te voeren die met de beide helften correspondeert. Desondanks is het voor de genoemde proefopzetten niet zinvol om de residuen van een hoofdeffectenmodel te onderzoeken.

4. Het werkt voor een gebruiker van gepubliceerde proefopzetten verwarrend dat de woordlengtepatronen voor minimum aberratie opzetten met 64 runs en 13, 14 of 15 factoren van Huang *et al*. (1998) meer aberratie vertonen dan de woordlengtepatronen voor de corresponderende opzetten van Chen *et al*. (1993).

   Chen, J., Sun, D.X., and Wu, C.F.J. (1993) A catalogue of two-level and three-level fractional factorial designs with small runs. *International Statistical Review* **61**: 131-145.
   Huang, P., Chen, D., and Voelkel, J.O. (1998): Minimum-aberration two-level split-plot designs. *Technometrics* **40**: 314-326.

5. Nelder and Lee (1998) en Lee and Nelder (1998) zetten hun lezers op het verkeerde been, doordat hun beschrijving van een procedure voor een gezamenlijke modellering van locatie en spreiding niet correspondeert met de procedure die zij in hun voorbeelden toepassen.

   Lee, Y., and Nelder, J.A. (1998) Generalized linear models for the analysis of quality-improvement experiments. *The Canadian Journal of Statistics* **26**: 95-105.
   Nelder, J.A., and Lee, Y. (1998). Joint modelling of mean and dispersion (letter to the editor). *Technometrics* **40**: 168-171.

6. De computersimulaties die Wolfinger en Tobias (1998) geven van een procedure om toegevoegde variantiecomponenten te modelleren als functie van variabelen die in een experiment worden gevarieerd geven ten onrechte geen informatie over de nulverdeling van toetsingsgrootheden voor de regressiecoëfficiënten in het model voor de genoemde variantiecomponenten.

   Wolfinger, R.D., and Tobias., R.D. (1998) Joint estimation of location, dispersion, and random effects in robust design. *Technometrics* **40**: 62-71.

7. Er is nader onderzoek gewenst van de volgende procedure om de toegevoegde variantiecomponent in een split-plot experiment te modelleren als functie van onafhankelijke variabelen. Laat *i* de whole plots indiceren en *j* de sub-plots. Laat de sub-plot variantie σ^2_{0ij} via een log-lineair model afhangen van zowel de variabelen die tussen whole plots worden gevarieerd als de variabelen die binnen whole plots worden gevarieerd. Laat de toegevoegde variantiecomponent tussen whole-plots σ^2_{1i} via een log-lineair model afhangen van de variabelen die tussen de whole plots worden gevarieerd. Dan kunnen de parameters van het model voor σ^2_{1i} geschat worden met een gegeneraliseerd lineair model voor de gekwadrateerde residuen van het locatiemodel voor de whole-plot gemiddelden met een gamma error en met de linkfunctie

$$\eta_i = \log\left(\Phi_i - A_i\right) = \log \sigma^2_{1i}$$

   waarbij Φ_i de verwachtingswaarde van het gekwadrateerde residu is en A_i het gefitte gemiddelde van de sub-plot variantie voor whole plot *i*. Het gefitte gemiddelde wordt verkregen uit de modellering van σ^2_{0ij} met het algoritme dat in Studie 3.3 van dit proefschrift beschreven staat.

8. De hoeveelheid muziek waaraan een stadsbewoner dagelijks ongevraagd wordt blootgesteld is om stil van te worden.