# Multi-phase experiments: from design to analysis
## Session 3: Multi-phase experiments

R. A. Bailey

University of St Andrews

QMUL (emerita)


Australasian Region of the International Biometric Society,
November 2015

## An experiment with two phases: expectation

first-phase design  second-phase design

set $\Gamma$ of treatments $\xleftarrow{h}$ set $\Omega_1$ of first-phase units $\xleftarrow{f}$ set $\Omega_2$ of second-phase units

random permutation of $\Omega_1$ | random permutation of $\Omega_2$

$$V_\Gamma \quad \leq \quad V_{\Omega_1} \quad \leq \quad V_{\Omega_2}$$

Expectation models $\leq V_\Gamma \leq V_{\Omega_1} \leq V_{\Omega_2}$



## An experiment with two phases: variance-covariance

first-phase design  second-phase design

set $\Gamma$ of treatments $\xleftarrow{h}$ set $\Omega_1$ of first-phase units $\xleftarrow{f}$ set $\Omega_2$ of second-phase units

random permutation of $\Omega_1$ | random permutation of $\Omega_2$

$\mathbf{V}_1 = \sum_i \eta_i \mathbf{Q}_i$  $\mathbf{V}_2 = \sum_j \xi_j \mathbf{P}_j$

Suppose that $f$ is equireplicate, with replication $r$.
On $\Omega_2$, the covariance matrix $\mathbf{V}_1$ becomes

$$\mathbf{V}_1 \otimes \mathbf{J}_r = \sum_i r\eta_i \left( \mathbf{Q}_i \otimes \frac{\mathbf{J}_r}{r} \right) = \sum_i r\eta_i \tilde{\mathbf{Q}}_i.$$

So the variance-covariance matrix for responses on second-phase units is $\sum_i r\eta_i \tilde{\mathbf{Q}}_i + \sum_j \xi_j \mathbf{P}_j$.



## Structure balance in the second phase

The variance-covariance matrix for responses on second-phase units is $\sum_i r\eta_i \tilde{\mathbf{Q}}_i + \sum_j \xi_j \mathbf{P}_j$.

If the second-phase design is orthogonal, the subspace of $V_{\Omega_2}$ with projector $\mathbf{P}_j$ is the direct sum of the following subspaces:

- the space $W_i$ corresponding to $\tilde{\mathbf{Q}}_i$, for all $i$ with $\tilde{\mathbf{Q}}_i \mathbf{P}_j \neq \mathbf{0}$;
- the residual subspace, which is the orthogonal complement of $V_{\Omega_1}$ in the $\mathbf{P}_j$-subspace.

Under structure balance, the first of these is replaced as follows:
if the c.e.f. for $W_i$ in the $\mathbf{P}_j$-subspace is $\lambda_{ij} \neq 0$,
use the subspace with projector $\lambda_{ij}^{-1} \mathbf{P}_j \tilde{\mathbf{Q}}_i \mathbf{P}_j$.



## Structure balance in both phases

| vector space | $V_\Gamma$ | $V_{\Omega_1}$ | $V_{\Omega_2}$ |
|---|---|---|---|
| orthogonal decomposition | $\mathbf{R}_k$ in $\mathcal{R}$ | $\mathbf{Q}_i$ in $\mathcal{Q}$ | $\mathbf{P}_j$ in $\mathcal{P}$ |

If the first-phase design is also structure-balanced,
then each subspace $W_i$ is decomposed in a similar way
by the treatment subspaces.

The good news is that the two sequential decompositions

($\mathbf{P}_j$-space by $\mathbf{Q}_i$ spaces) by $\mathbf{R}_k$-spaces

and

$\mathbf{P}_j$-space by ($\mathbf{Q}_i$ spaces by $\mathbf{R}_k$-spaces)

are the same,
and the canonical efficiency factors just multiply.



## Meatloaves: a tasting experiment, Phase I

2 Rosemary
3 Irradiation

3 Blocks
6 Loaves in B

3 Sessions
12 Panellists in S
6 Time-orders in S

6 treatments  18 meatloaves  216 tastings

First-phase design is a complete-block design.
It is randomized by randomly permuting blocks and
randomly permuting loaves within each block.
This gives the covariance model

$$\gamma_{\mathrm{BL}}\mathbf{I}_{\mathrm{L}} + \gamma_{\mathrm{B}}(\mathbf{J}_{\mathrm{B}} - \mathbf{I}_{\mathrm{L}}) + \gamma_0(\mathbf{J} - \mathbf{J}_{\mathrm{B}}) = \eta_{\mathrm{BL}}\mathbf{Q}_{\mathrm{BL}} + \eta_{\mathrm{B}}\mathbf{Q}_{\mathrm{B}} + \eta_0\mathbf{Q}_0$$

$\uparrow \quad \uparrow \quad \uparrow$ non-negative



## Meatloaves: a tasting experiment, Phase II

2 Rosemary
3 Irradiation

3 Blocks
6 Loaves in B

3 Sessions
12 Panellists in S
6 Time-orders in S

6 treatments  18 meatloaves  216 tastings

The second-phase design allocates blocks to sessions and
uses a pair of $6 \times 6$ Latin squares for the loaves in each block.
It is randomized by randomly permuting sessions,
randomly permuting panellists within each session,
and randomly permuting time-orders within each session.
This gives the covariance model

$$\xi_0\mathbf{P}_0 + \xi_{\mathrm{S}}\mathbf{P}_{\mathrm{S}} + \xi_{\mathrm{SP}}\mathbf{P}_{\mathrm{SP}} + \xi_{\mathrm{ST}}\mathbf{P}_{\mathrm{ST}} + \xi_{\mathrm{SPT}}\mathbf{P}_{\mathrm{SPT}}.$$

Because the second-phase design has equal replication 12,
the overall covariance matrix for the 216 responses is

$$\xi_0\mathbf{P}_0 + \xi_{\mathrm{S}}\mathbf{P}_{\mathrm{S}} + \xi_{\mathrm{SP}}\mathbf{P}_{\mathrm{SP}} + \xi_{\mathrm{ST}}\mathbf{P}_{\mathrm{ST}} + \xi_{\mathrm{SPT}}\mathbf{P}_{\mathrm{SPT}} + 12\eta_0\tilde{\mathbf{Q}}_0 + 12\eta_{\mathrm{B}}\tilde{\mathbf{Q}}_{\mathrm{B}} + 12\eta_{\mathrm{BL}}\tilde{\mathbf{Q}}_{\mathrm{BL}}$$

## Meatloaves: skeleton anova

2 Rosemary
3 Irradiation

3 Blocks
6 Loaves in B

3 Sessions
12 Panellists in S
6 Time-orders in S

6 treatments | 18 meatloaves | 216 tastings

| tastings | | meatloaves | | treatments | | |
|---|---|---|---|---|---|---|
| source | df | source | df | source | df | EMS |
| Mean | 1 | Mean | 1 | Mean | 1 | $\xi_0 + 12\eta_0 + q_0$ |
| Sessions | 2 | Blocks | 2 | | | $\xi_S + 12\eta_B$ |
| Panellists[S] | 33 | | | | | $\xi_{SP}$ |
| Time-orders[S] | 15 | | | | | $\xi_{ST}$ |
| P#T[S] | 165 | Loaves[B] | 15 | Rosemary | 1 | $\xi_{SPT} + 12\eta_{BL} + q(R)$ |
| | | | | Irradiation | 2 | $\xi_{SPT} + 12\eta_{BL} + q(I)$ |
| | | | | R# I | 2 | $\xi_{SPT} + 12\eta_{BL} + q(RI)$ |
| | | | | Residual | 10 | $\xi_{SPT} + 12\eta_{BL}$ |
| | | Residual | 150 | | | $\xi_{SPT}$ |



## Residual degrees of freedom

Lesson

*If treatments are applied in Phase I,*
*the number of degrees of freedom for the relevant residual*
*cannot increase in Phase II.*

Principle

*If treatments are orthogonal to 'large blocks' in Phase I,*
*then those large blocks should be confounded with*
*"large blocks" in Phase II.*



## Estimating spectral components of variance

In the meatloaves example,
there is one EMS equal to $\xi_S + 12\eta_B$, and neither of those spectral components occurs in any other EMS.

Potential Problem

*If an $\eta$-component and a $\xi$-component occur together but nowhere else, then neither can be estimated.*
*This may not matter.*

Spectral components of variance must be non-negative. In a single-phase experiment, each spectral component is estimated by the relevant residual mean square, which is non-negative. In the meatloaves example, there is one residual EMS equal to $\xi_{SPT} + 12\eta_{BL}$ and another equal to $\xi_{SPT}$.

Potential Problem

*In a two-phase experiment,*
*equating actual residual mean squares to their expections may lead to negative estimates for spectral components.*



## Simple atheletes

3 Intensities
3 Surfaces

4 Months
3 Athletes in M
3 Tests in A, M

4 Batches
9 Locations in B

The first-phase design allocates intensities to athletes within each month, and allocates surfaces to tests within each athlete.

Treatments are orthogonal to Months,
so the second-phase design allocates months to batches.

The upcoming anova assumes appropriate randomization in both phases.



## Skeleton anova for simple athletes

| locations | | tests | | conditions | | |
|---|---|---|---|---|---|---|
| source | df | source | df | source | df | EMS |
| Mean | 1 | Mean | 1 | Mean | 1 | $\xi_0 + 12\eta_0 + q_0$ |
| Batches | 3 | Months | 3 | | | $\xi_B + \eta_M$ |
| Locations[B] | 32 | Athletes[M] | 8 | Intensities | 2 | $\xi_{BL} + \eta_{MA} + q(I)$ |
| | | | | Residual | 6 | $\xi_{BL} + \eta_{MA}$ |
| | | Tests[A[M]] | 24 | Surfaces | 2 | $\xi_{BL} + \eta_{MAT} + q(S)$ |
| | | | | I#S | 4 | $\xi_{BL} + \eta_{MAT} + q(I\#S)$ |
| | | | | Residual | 18 | $\xi_{BL} + \eta_{MAT}$ |



## Same number of units in both phases

Suppose that the design for the second phase has replication $r$.
What happens when $r = 1$?

Lesson

*If $r = 1$, the only way to achieve structure balance is to have an orthogonal design.*

Potential Problem

*If $r = 1$ then every EMS contains one $\xi$-component and one $\eta$-component, so it is impossible to estimate any spectral component. However, it may be possible to estimate differences such as $\eta_{MA} - \eta_{MAT}$.*

## Wheat experiment

49 Varieties

4 Blocks
49 Plots in B

4 Intervals
7 Runs in I
7 Times in I

The design in the first phase is a complete-block design.
This is orthogonal.

The design in the second phase has replication 1.
Blocks will be allocated to intervals.
To get structure balance, we need to make it orthogonal, so we need to introduce pseudofactors $P_R$ and $P_T$ for Plots in Blocks, such that $P_R$ is confounded with Runs in each interval and $P_T$ is confounded with Times in each Interval.



## Pseudofactors in the wheat experiment

49 Varieties

4 Blocks
7 $P_R$ in B
7 $P_T$ in B

4 Intervals
7 Runs in I
7 Times in I

Write the 49 varieties in a $7 \times 7$ square array; write out a complete set of six mutually orthogonal $7 \times 7$ Latin squares.

| Block | $P_R$ | $P_T$ |
|---|---|---|
| 1 | array rows | array columns |
| 2 | letters of LS 1 | letters of LS 2 |
| 3 | letters of LS 3 | letters of LS 4 |
| 2 | letters of LS 5 | letters of LS 6 |

Now the design for the first phase is no longer orthogonal!



## Skeleton anova for wheat experiment

49 Varieties

4 Blocks
7 $P_R$ in B
7 $P_T$ in B

4 Intervals
7 Runs in I
7 Times in I

| analyses | | plots | | varieties | | | |
|---|---|---|---|---|---|---|---|
| source | df | source | df | cef | source | df | EMS |
| Mean | 1 | Mean | 1 | 1 | Mean | 1 | $\xi_0 + \eta_0 + q_0$ |
| Intervals | 3 | Blocks | 3 | | | | $\xi_I + \eta_B$ |
| Runs[I] | 24 | $P_R$[B] | 24 | 1/4 | $T_1$ | 24 | $\xi_{IR} + \eta_{BP} + \frac{1}{4}q(T_1)$ |
| Times[I] | 24 | $P_T$[B] | 24 | 1/4 | $T_2$ | 24 | $\xi_{IT} + \eta_{BP} + \frac{1}{4}q(T_2)$ |
| R#T[I] | 144 | P[B]$\vdash P_T \vdash P_R$ | 144 | 3/4 | $T_1$ | 24 | $\xi_{IRT} + \eta_{BP} + \frac{3}{4}q(T_1)$ |
| | | | | 3/4 | $T_2$ | 24 | $\xi_{IRT} + \eta_{BP} + \frac{3}{4}q(T_2)$ |
| | | | | | Resl | 96 | $\xi_{IRT} + \eta_{BP}$ |



## Pseudofactors for the units in the first phase: I

Lesson

*Sometime we need to introduce pseudofactors for the units in the first phase, in order to make the design for the second phase orthogonal (or even structure-balanced).*

Potential Problem

*Even if the design for the first phase is originally orthogonal, it may no longer be so once there are pseudofactors on the second-phase units, and we need to be aware of this in both the design and the analysis of the experiment.*



## Pseudofactors for the units in the first phase: II

Potential Problem

*Different pseudofactors for the same real factor have the same spectral component of variance. Hence it is possible to have residual mean squares whose EMS are linearly dependent (this did not happen in the wheat example). Then there is no uniformly best unbiased quadratic estimator of the spectral components of variance involved, even those that are estimable.*

Potential Problem

*The introduction of pseudofactors on the first-phase units may make that structure something other than a poset block structure (this did not happen in the wheat example). So we need to be able to think about structures, in both the design and analysis, without being on auto-pilot for poset block structures.*



## Hard atheletes

3 Intensities
3 Surfaces

4 Months
3 Athletes in M
3 Tests in A, M

4 Batches
9 Locations

Locations are now crossed with Batches.

There are 36 units in the first phase and 36 units in the second phase, so the design for the second phase needs to be orthogonal.

There are 12 athletes and 9 locations, and $12 \times 9 > 36$, so how can this be achieved?

## Making Months and Athletes both orthogonal to Locations

| Month | Athlete | Location | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 1 | • | • | • | | | | | | |
| 2 | 4 | • | • | • | | | | | | |
| 3 | 7 | • | • | • | | | | | | |
| 4 | 10 | • | • | • | | | | | | |
| 1 | 2 | | | | • | • | • | | | |
| 2 | 5 | | | | • | • | • | | | |
| 3 | 8 | | | | • | • | • | | | |
| 4 | 11 | | | | • | • | • | | | |
| 1 | 3 | | | | | | | • | • | • |
| 2 | 6 | | | | | | | • | • | • |
| 3 | 9 | | | | | | | • | • | • |
| 4 | 12 | | | | | | | • | • | • |

This gives a factor $A \vee L$ with three levels.



## Supremum of two factors

Given two factors $A$ and $L$, the levels of the factor $A : L$ are all combinations of the levels of $A$ with the levels of $L$.

I usually write this factor as $A \wedge L$, because it is the infimum of $A$ and $L$.

This means that

- $A$ and $L$ are both marginal to $A \wedge L$;
- subject to this, $A \wedge L$ has the least number of levels.

The supremum $A \vee L$ of $A$ and $L$ satisfies

- $A \vee L$ is marginal to both $A$ and $L$;
- subject to this, $A \vee L$ has the most number of levels.

If the software does not recognize a non-trivial supremum like $A \vee L$ and fit it separately, then it will be included in whichever is fitted first of $A$ and $L$.

This can cause problems even in single-phase experiments.



## Hard atheletes: a pseudofactor

3 Intensities
3 Surfaces

4 Months
3 P
3 Tests in P, M

4 Batches
9 Locations

As a pseudofactor, $P$ is crossed with Months,
Tests is nested in both, and $M \wedge P = A$.
We will use a design that ensures that $P = A \vee L$.

Then Tests[A[M]] is split into two orthogonal parts:
$T_L$ is the part of Locations orthogonal to $P$;
$T_{BL}$ is the part of B#L orthogonal to Athletes.

How do we choose the design for the first phase?
If we allocate Intensities to levels of $P$ then Intensities are confounded with Locations as well as with Athletes, so will probably have an unnecessarily high variance.
Otherwise, the design for the first phase must be non-orthogonal.



## Hard athletes: anova for chosen design

| locations | | tests | | conditions | | | |
|---|---|---|---|---|---|---|---|
| source | df | source | df | cef | source | df | EMS |
| Mean | 1 | Mean | 1 | 1 | Mean | 1 | $\xi_0 + \eta_0 + q_0$ |
| Batches | 3 | Months | 3 | | | | $\xi_B + \eta_M$ |
| Locations | 8 | P | 2 | 1/16 | Intensities | 2 | $\xi_L + \eta_{MA} + \frac{1}{16}q(I)$ |
| | | $T_L$ | 6 | 1/16 | Surfaces | 2 | $\xi_L + \eta_{MAT} + \frac{1}{16}q(S)$ |
| | | | | 1/4 | I#S | 4 | $\xi_L + \eta_{MAT} + \frac{1}{4}q(I\#S)$ |
| B#L | 24 | M:P | 6 | 15/16 | Intensities | 2 | $\xi_{BL} + \eta_{MA} + \frac{15}{16}q(I)$ |
| | | | | | Residual | 4 | $\xi_{BL} + \eta_{MA}$ |
| | | $T_{BL}$ | 18 | 15/16 | Surfaces | 2 | $\xi_{BL} + \eta_{MAT} + \frac{15}{16}q(S)$ |
| | | | | 3/4 | I#S | 4 | $\xi_{BL} + \eta_{MAT} + \frac{3}{4}q(I\#S)$ |
| | | | | | Residual | 12 | $\xi_{BL} + \eta_{MAT}$ |



## Proteomics

2 Interventions
2 Tissues

8 Cages
2 Animals in C
2 Positions in A, C

8 Runs
8 Labels

In the first-phase design, Interventions are applied to whole Cages; two different types of tissue are taken from each Animal.

We expect that $\eta_C > \eta_{CA} > \eta_{CAP}$ and $\xi_R > \xi_{RL}$ and $\xi_L > \xi_{RL}$.

Interventions has the largest variance from the first phase, so we should try to confound it with the lowest-variance source R#L in the second phase.

In the first phase, there are only 6 residual df in Cages.
In order not to reduce the relevant residual df for Interventions, we should try to confound the whole of Cages with the same source in the second phase.

Also Tissues and I#T should be orthogonal to Runs and Labels.



## Factors which are 'hard to set' or which must be applied to large areas in Phase I

### Lesson

*If a treatment factor is 'hard to set' or a 'main-plot factor' in Phase I, then it is probably in a Phase I stratum with large variance. Stratum variances from the two phases are added.*

### Principle

*If a treatment factor is 'hard to set' or a 'main-plot factor in Phase I, then it should be allocated to a Phase II stratum with small variance.*

### Principle

*If a treatment factor is 'hard to set' or a 'main-plot factor in Phase I, then its Phase I stratum should not be split in the second-phase design, to avoid reducing residual degrees of freedom.*

## Tobacco mosaic virus

4 Treatments

2 Sets
4 Nicotine plants in S
4 Positions

4 Lots
4 Datura plants in L
4 Assay positions
2 Halfleaves in A, D, L

The design in the first phase uses a Latin square in each Set.

The design in the second phase uses a 4-level pseudofactor Pairs on the nicotine plants. It is crossed with Sets.
The pairs are allocated to the Lots.

Within each pair,
the 8 combinations of 2 nicotine plants with 4 positions are allocated to the halfleaves in 4 datura plants $\times$ 4 assay positions using a semi-Latin square: the difference between nicotine plants is orthogonal to rows, columns and whole leaves; the other 6 d.f. have cef 1/2 in halfleaves and cef 1/2 in whole leaves.



| datura half-leaves | | tobacco leaves | | | treatments | | |
|---|---|---|---|---|---|---|---|
| source | df | cef | source | df | cef | source | df |
| Mean | 1 | 1 | Mean | 1 | 1 | Mean | 1 |
| Lots | 3 | 1 | Pairs | 3 | | | |
| Assay Posn | 3 | | | | | | |
| Datura[L] | 12 | | | | | | |
| A#L | 9 | | | | | | |
| A#D[L] | 36 | 1/2 | Posn | 3 | | | |
| | | 1/2 | Posn#Set | 3 | | | |
| | | 1/2 | Posn#Nic[S] | 18 | 1/2 | Tmt | 3 |
| | | | | | 1/2 | Residual | 15 |
| | | 1 | Residual | 12 | | | |
| Half[A#D[L]] | 64 | 1 | Set | 1 | | | |
| | | 1 | Set#Pair | 3 | | | |
| | | 1/2 | Posn | 3 | | | |
| | | 1/2 | Posn#Set | 3 | | | |
| | | 1/2 | Posn#Nic[S] | 18 | 1/2 | Tmt | 3 |
| | | | | | 1/2 | Residual | 15 |
| | | 1 | Residual | 36 | | | |



## Non-orthogonality in the second phase

The variance-covariance matrix of the responses on the units in the second phase is

$$\mathbf{V} = \sum_i \eta_i \mathbf{Q}_i + \sum_j \xi_j \mathbf{P}_j,$$

where the idempotent matrices $\mathbf{Q}_i$ and $\mathbf{P}_j$ are known.

If the design in the second phase is not orthogonal, then the $\mathbf{Q}_i$ do not all commute with all of the $\mathbf{P}_j$, and so we do not know the eigenspaces of $\mathbf{V}$.

Consequently, ANOVA does not give the best estimators.

The skeleton anova table is still a useful summary of the properties of the design.
Such tables can be used at the design stage to compare different potential designs.



## Treatment factors in the second phase

Sometimes, some treatment factors are applied only in the second phase.

### Principle

*Plan the whole thing in advance, especially if there may be interactions between the first-phase treatments and the second-phase treatments.*

### Principle

*If second-phase treatments are orthogonal to 'large blocks' in Phase II, then those large blocks should be confounded with "large blocks" in Phase I.*

### Principle

*If a treatment factor is 'hard to set' or a 'main-plot factor in Phase II, then try to confound it with a Phase I stratum with small variance, and try to avoid further reducing its residual degrees of freedom.*



## 'Big with big' or not?

If treatments in either phase can be orthogonal to blocks in that phase, then confound big with big to leave more residual degrees of freedom.

If treatments in either phase must be allocated to large units in that phase, then try to make these large units orthogonal to blocks in the other phase.

If the overall design is a fractional factorial in the treatment factors, then gaining residual df for hard-to-set factors may be at the expense of losing residual df for other treatment factors. A compromise may be needed whereby blocks in the two phases are deliberately given a supremum with a small number of levels.

If the numbers are such that treatments cannot be orthogonal to blocks in either phase, then the best strategy depends on the possible cefs in each phase and on the (unknown) relative sizes of all of the spectral components of variance involved.



## References


- C. J. Brien and R. A. Bailey: Multiple randomizations. *J. Roy. Statist. Soc. Ser. B* **68** (2006), 571–609.
- C. J. Brien and R. A. Bailey: Decomposition tables for experiments. I. A chain of randomizations. *Ann. Statist.* **37** (2009), 4184–4213.
- C. J. Brien, B. D. Harch, R. L. Correll and R. A. Bailey: Multiphase experiments with at least one later laboratory phase. I. Orthogonal designs. *J. Agric. Biol. Env. Statist.* **16** (2011), 422–450.
- R. A. Bailey and C. J. Brien: Randomization-based models for multitiered experiments: I. A chain of randomizations. *Ann. Statist.*, in press.
- R. A. Bailey: Structures defined by factors. In *Handbook of Design and Analysis of Experiments* (eds. A. Dean, M. Morris, J. Stufken and D. Bingham), Chapman and Hall/CRC (2015), pp. 371–414.