The design and statistical analysis of experiments involving laboratory animals

Michael FW Festing, Ph.D., D.Sc., CStat.

michaelfesting@aol.com

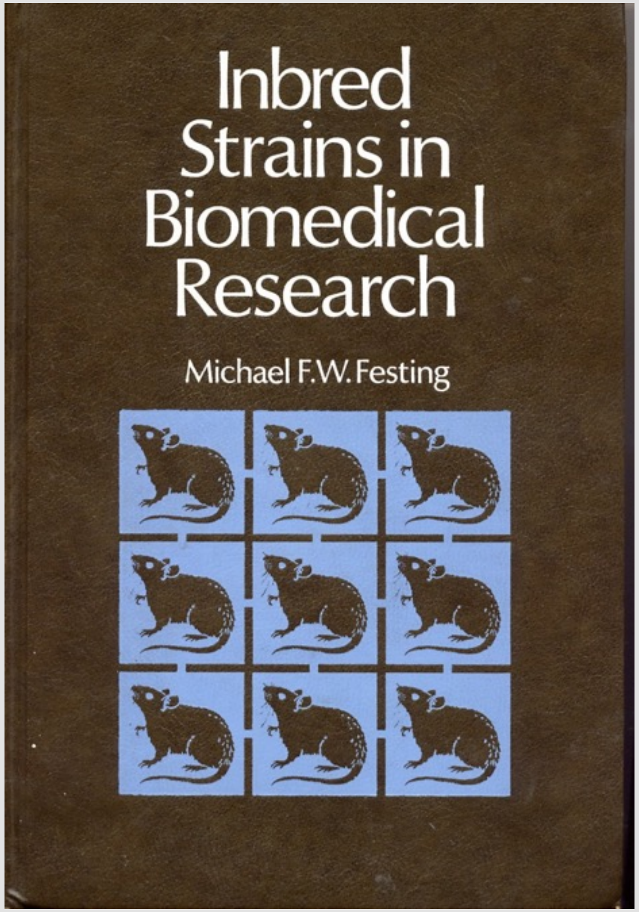
www.3Rs-reduction.co.uk

A PPL course

Michael FW Festing, Ph.D., D.Sc, Cstat

1966-1981 Geneticist, MRC Laboratory animals centre

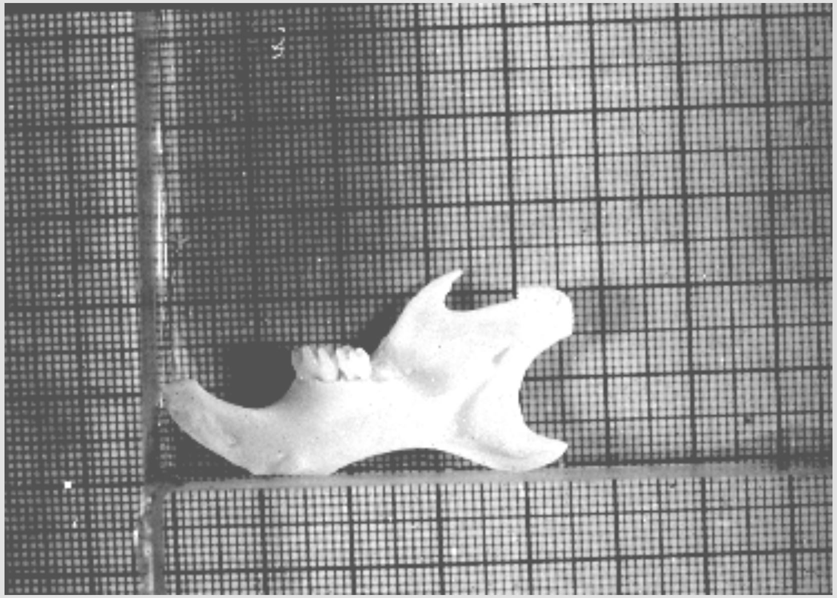
Aim of the LAC: To supply high quality, disease-free breeding stock to research workers and commercial breeders.

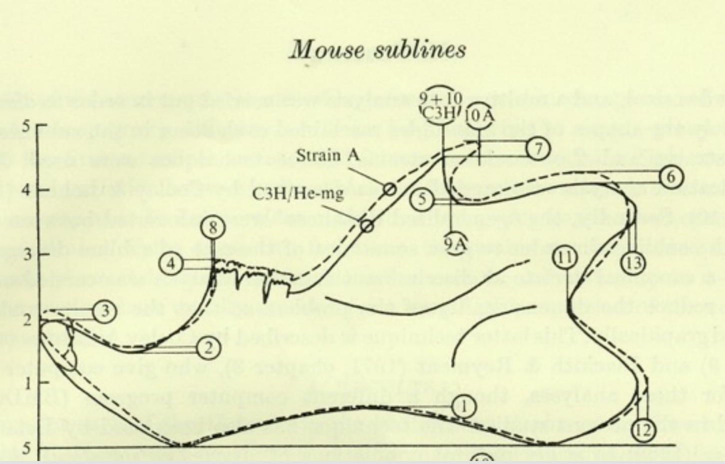


Inbred Strains in Biomedical Research

Michael F.W. Festing

Some personal research: Mandible shape for genetic quality control c1970s





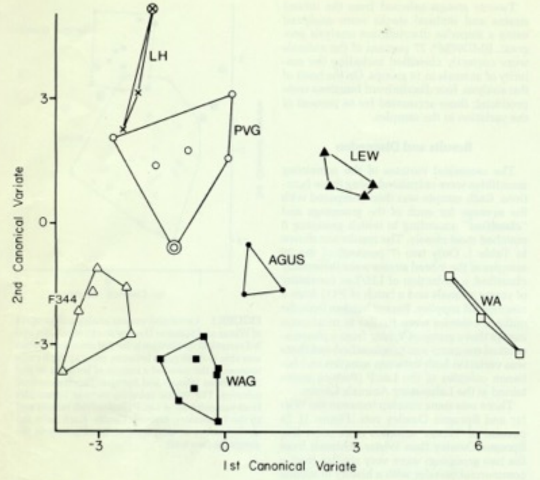


FIGURE 2 Canonical variate analysis of inbred strains of rats; 1st and 2nd canonical variates. Each point represents mean of a sample of between 8 and 38 rats. Two colonies, one LH and PVG (encircled), were poor fits against their group means (see text). Four samples of LEW/SsN rats were obtained from a colony held at the Laboratory Animals Centre in 1978/79 and prepared in a similar manner. (Samples of the other inbred strains maintained at the LAC sampled in 1977/79 fitted well.)

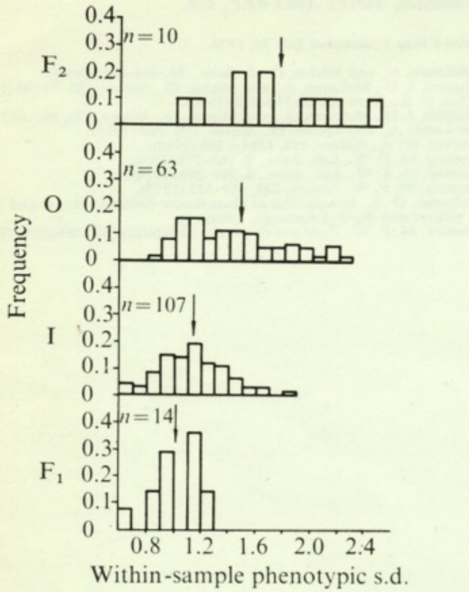


Fig. 1 Distribution of the within-sample phenotypic standard deviation of an index of mandible shape in F_2 hybrid, outbred (O), inbred (I) and F_1 hybrid mice. Arrows indicate the mean, n is the number of samples. Sample size averaged 12 mice.

Some personal research: Strain differences in escape time in a water maze

Water Escape Learning in Mice. I

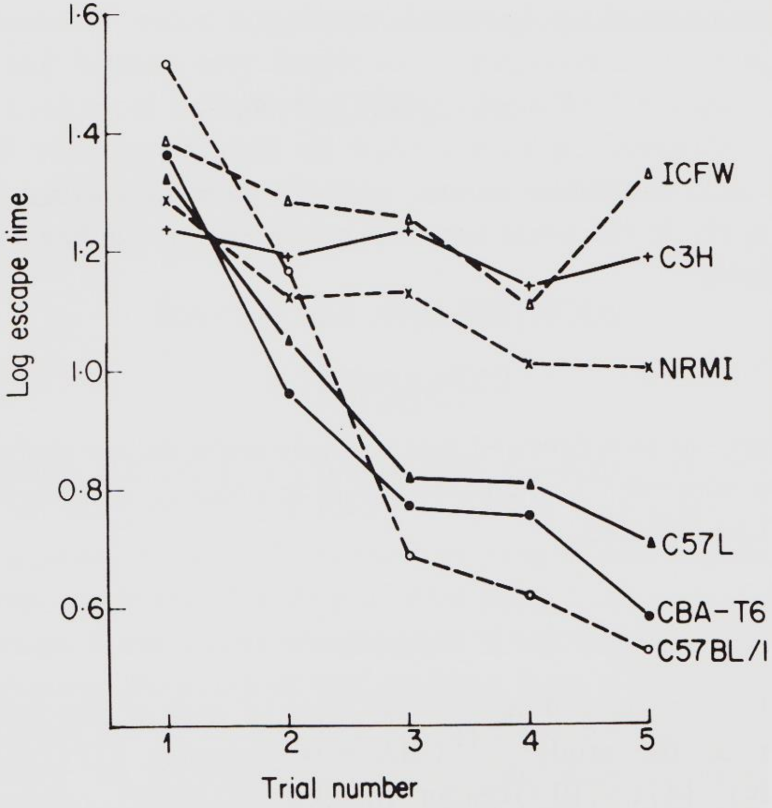


Fig. 1. Strain means for experiment 1.

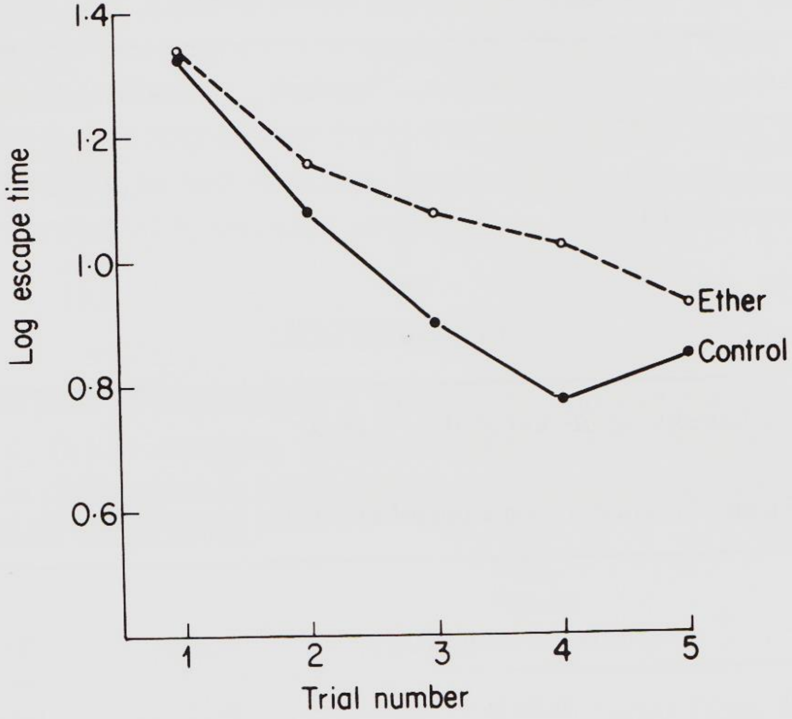


Fig. 2. Learning curves averaged across strains.

Some personal research: Exercise in a running wheel

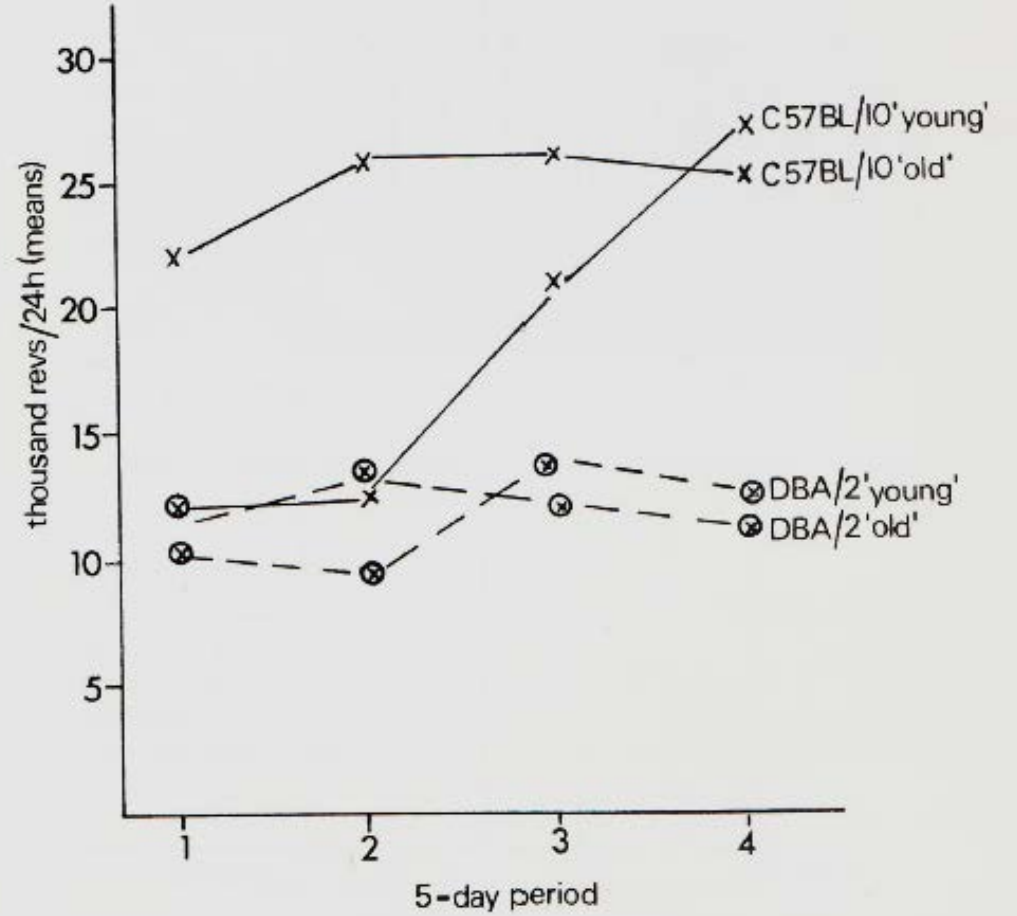


Fig. 3. Genetic and age differences in wheel running by mice. Each point represents the mean number of revolutions of 2 mice over a 5-day period.

The design and statistical analysis of experiments involving laboratory animals

Principles of Humane Experimental Technique

Russell and Burch 1959

- ❑ Replacement
  - ❑ e.g. in-vitro methods, less sentient animals
- ❑ Refinement
  - ❑ e.g. anaesthesia and analgesia, environmental enrichment
- ❑ Reduction
  - ❑ Research strategy
    - ❑ Shotgun vs Fundamental
  - ❑ Controlling variability
    - ❑ Genetics, appropriate model
    - ❑ (disease)
  - ❑ Experimental design and statistics

FRAME

Concern about the quality of animal research expressed in 1992

Outlined the principles of good experimental design and did a small survey of published papers (mostly toxicology)

1. Few used randomised block designs even though this is the most common design in agricultural and industrial research.

2. Factorial designs rare although they provide extra information at no extra cost

Festing, M. F. W. "The scope for improving the design of laboratory animal experiments." Laboratory Animals 26 (1992): 256-67.

Won first prize in a GV-SOLAS competition for the best published or unpublished paper on laboratory animal science

Concern about the quality of animal research

A meta-analysis of 44 randomised controlled animal studies of fluid resuscitation

- Only 2 said how animals had been allocated
- None had sufficient power to detect reliably a halving in risk of death
- Substantial scope for bias
- Substantial heterogeneity in results, due to method of inducing the bleeding
- Odds ratios impossible to interpret
- Authors queried whether these animal experiments made any contribution to human medicine

Roberts et al 2002, BMJ 324:474

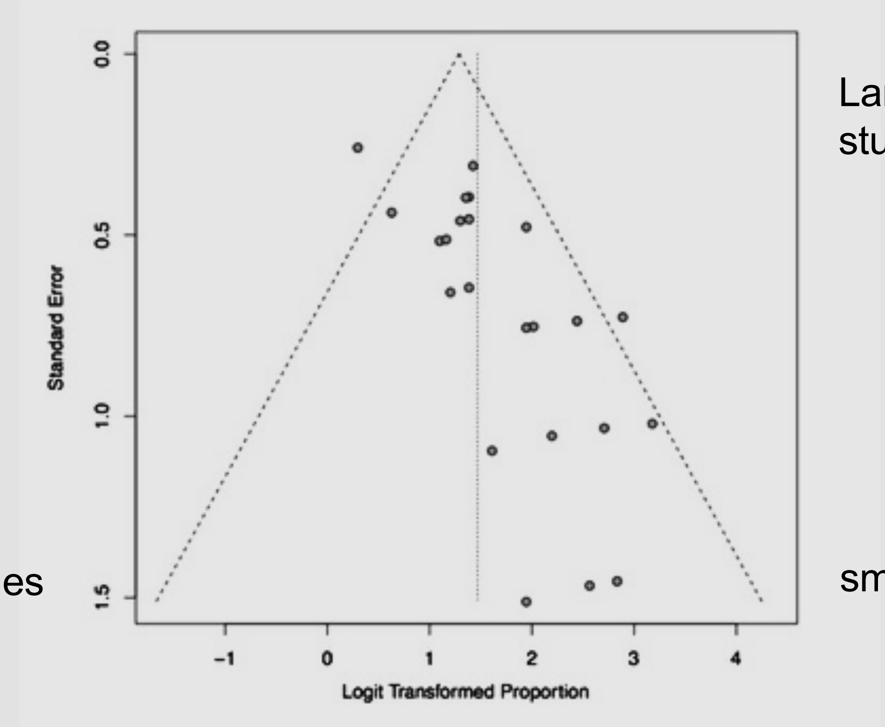
Six meta-analyses showing poor agreement between animal and human responses, 2007

Intervention	Human results	Animal results (meta-analysis)	Agree?
Corticosteroids for head injury	No improvement	Improved nurological outcome n=17	No
Antofibrinolytics for surgery	Reduces blood loss	Too little good quality data n=8	No
Thrombolysis with TPA for acute ischaemic stroke	Reduces death	Reduces death but publication bias and overstatement (n=113)	Yes
Tirilazad for stroke	Increases risk of death	Reduced infarct volume and improved behavioural score n=18	No
Corticosteroids for premature birth	Reduces mortality	Reduces mortality n=56	Yes
Bisphosphonates for osteoperosis	Increase bone density	Increase bone density n=16	Yes

Perel et al (2007) BMJ 334:197-200

Funnel plots and publication bias

Each dot is one experiment. Small negatives have remained unpublished.



Large powerful studies

small negative studies

small positive studies

Funnel plot demonstrating possible but not statistically significant publication bias in assessment of pain (*P* > 0.05). -Dashed diagonal lines indicate 95% CI

J Ther Ultrasound. 2017 Apr 1;5:9. doi: 10.1186/s40349-017-0080-4. eCollection 2017.
A meta-analysis of palliative treatment of pancreatic cancer with high intensity focused ultrasound.
Dababou S[1], Marrocchio C[1], Rosenberg J[2], Bitton R[2], Pauly KB[2], Napoli A[3], Hwang JH[4], Ghanouni P[2].

Problems with published papers

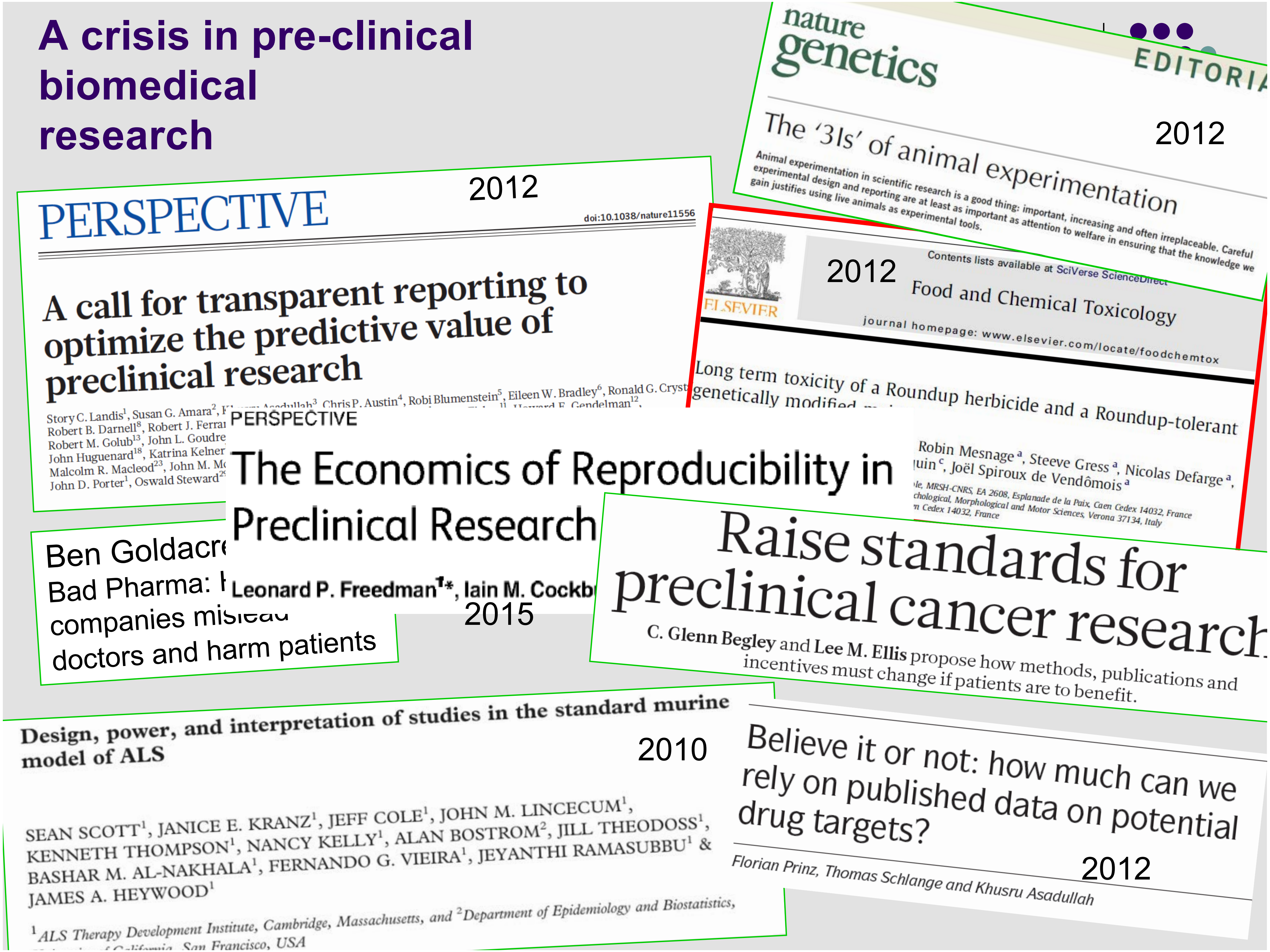
Survey of a random sample of 271 published papers using laboratory animals

Of the papers studied:

- 87% did not report random allocation of subjects to treatments
- 86% did not report "blinding" where it seemed to be appropriate
- 100% failed to justify the sample sizes used
- 5% did not clearly state the purpose of the study
- 6% did not indicate how many separate experiments were done
- 13% did not identify the experimental unit
- 26% failed to state the sex of the animals
- 24% reported neither age not weight of animals
- 4% did not mention the number of animals used
- 35% which reported numbers used these differed in the materials and methods and the results sections
- etc.

Kilkenny et al (2009), PLoS One Vol. 4, e7824

A crisis in pre-clinical biomedical research



nature genetics

EDITORIA

2012

The '3Is' of animal experimentation

Animal experimentation in scientific research is a good thing: important, increasing and often irreplaceable. Careful experimental design and reporting are at least as important as attention to welfare in ensuring that the knowledge we gain justifies using live animals as experimental tools.

2012

PERSPECTIVE

doi:10.1038/nature11556

A call for transparent reporting to optimize the predictive value of preclinical research

Story C. Landis[1], Susan G. Amara[2], K
Robert B. Darnell[8], Robert J. Ferra
Robert M. Golub[13], John L. Goudre
John Huguenard[18], Katrina Kelner
Malcolm R. Macleod[23], John M. M
John D. Porter[1], Oswald Steward[2]

Chris P. Austin[4], Robi Blumenstein[5], Eileen W. Bradley[6], Ronald G. Cryst
Howard E. Gendelman[12],

2012

ELSEVIER

Contents lists available at SciVerse ScienceDirect

Food and Chemical Toxicology

journal homepage: www.elsevier.com/locate/foodchemtox

Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified

Robin Mesnage[a], Steeve Gress[a], Nicolas Defarge[a],
uin[c], Joël Spiroux de Vendômois[a]

le, MRSH-CNRS, EA 2608, Esplanade de la Paix, Caen Cedex 14032, France
chological, Morphological and Motor Sciences, Verona 37134, Italy
n Cedex 14032, France

PERSPECTIVE

The Economics of Reproducibility in Preclinical Research

Leonard P. Freedman[1]*, Iain M. Cockb

2015

Ben Goldacre
Bad Pharma: H
companies mislead
doctors and harm patients

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Design, power, and interpretation of studies in the standard murine model of ALS

2010

SEAN SCOTT[1], JANICE E. KRANZ[1], JEFF COLE[1], JOHN M. LINCECUM[1], KENNETH THOMPSON[1], NANCY KELLY[1], ALAN BOSTROM[2], JILL THEODOSS[1], BASHAR M. AL-NAKHALA[1], FERNANDO G. VIEIRA[1], JEYANTHI RAMASUBBU[1] & JAMES A. HEYWOOD[1]

[1] ALS Therapy Development Institute, Cambridge, Massachusetts, and [2] Department of Epidemiology and Biostatistics, University of California, San Francisco, USA

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

2012

SOD1^{G93A}: The standard model for FALS and ALS

Scott et al (2008) Amyotrophic Lateral Sclerosis 9:4-15

- >50 papers describing therapeutic agents which extend lifespan in mice
- Only one (riluzole) has any clinical effect
- Scott et al:
  - Confounding factors (gender, litter, censoring, copy number) identified & controlled.
  - Power analysis used to determine an appropriate sample size
  - 70 compounds subsequently tested. None (including riluzole) increased survival.
- “The majority of published effects are most likely measurements of noise in the distribution of survival means as opposed to actual drug effects.”

Cost of irreproducible pre-clinical research in the USA alone

US$28,000,000,000 per annum (US$28 billion)

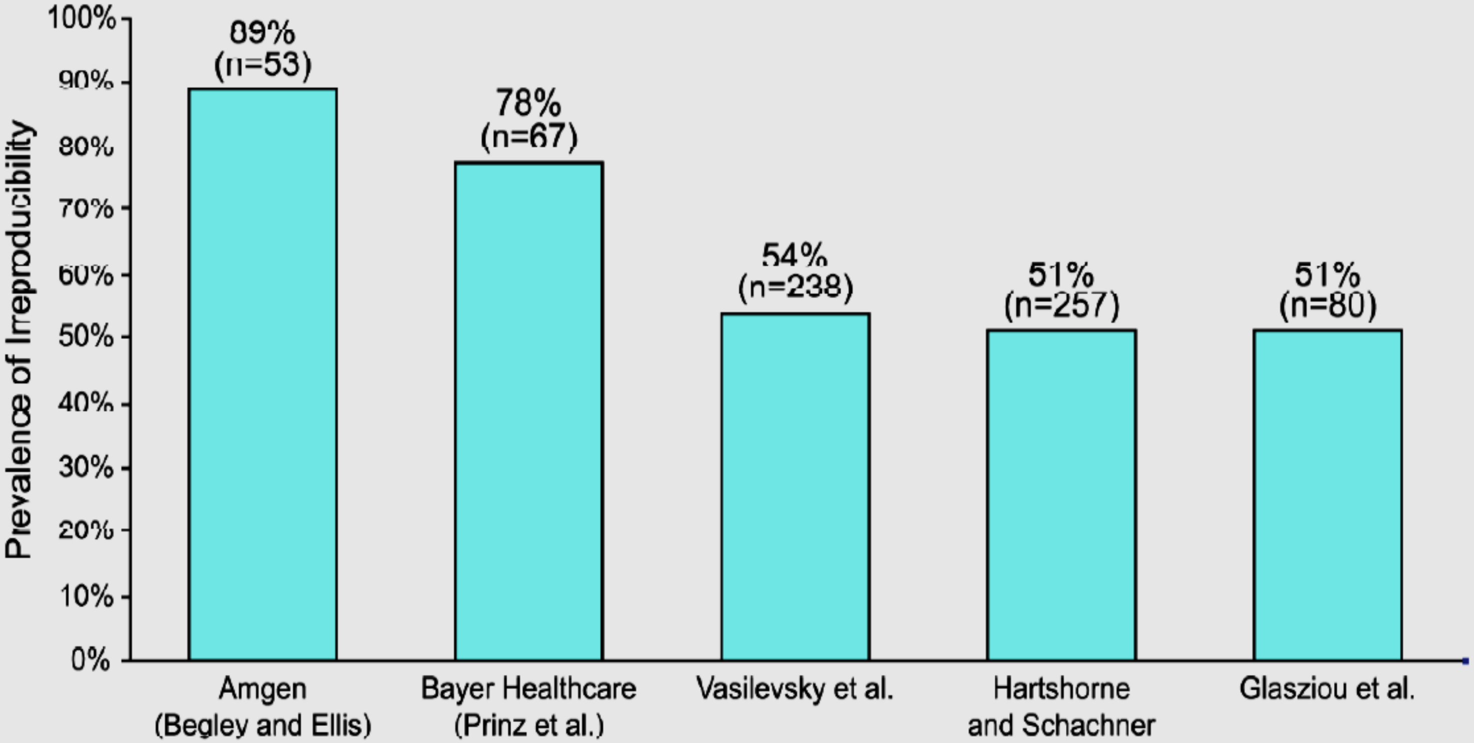


Fig 1. Studies reporting the prevalence of irreproducibility. Source: Begley and Ellis [6], Prinz et al. [7], Vasilevsky [8], Hartshorne and Schachner [5], and Glasziou et al. [9].

doi:10.1371/journal.pbio.1002165.g001

Freedman et al (2015)

Some possible causes of lack of repeatability (false positives)

- Bias: incorrect or no randomisation/blinding (Due to use of the "Completely randmized" experimental design).
- Pseudo-replication: failure to identify the experimental unit correctly with over-estimation of "n" (e.g. animals/cage)
- Wrong animals (large species/strain differences in mice and rats)
- Failure to repeat or build in repetition (e.g. using randomised block designs). (*In-vitro* experiments "repeat the experiment 3 times")
- Under-powered. Negative results remain unpublished. Excessive false positives due to the 5% significance level
- Technical errors. E.g. wrong monoclonal Abs.
- Statistical errors. E.g. assumptions invalid when doing parametric tests
- Fraud

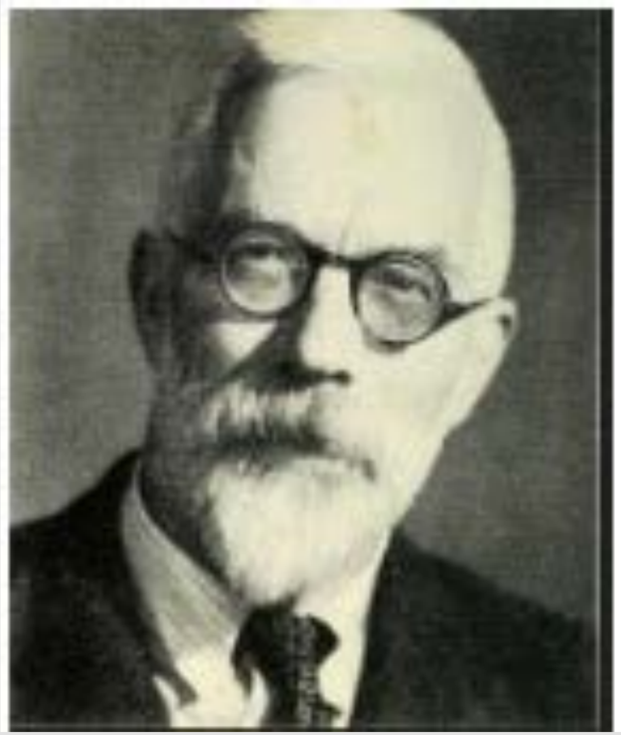
Clear evidence of conflicts of interest impacting results

Positive results in studies of endocrine disruption by bisphenol A.

94/104	= 90%	Government funded
0/11	= 0%	Industry funded

Frederick S. vom Saal and Claude Hughes.
***Environ Health Perspect* 113:926–933 (2005)**

The father of the randomised, controlled experiment



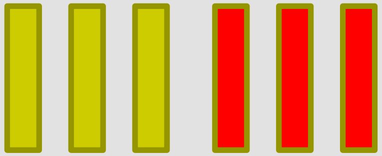
Sir Ronald Aylmer Fisher FRS (1890 – 1962), who published as R. A. Fisher, was an English statistician, and biologist, who used mathematics to combine Mendelian genetics and natural selection,... wikipedia.org

“To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.”

The randomised controlled experiment: basic principles

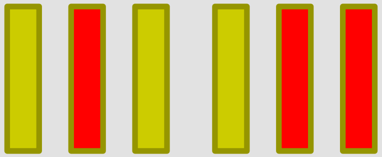
Developed at the Rothamsted Experimental Station in the 1920s, largely by RA Fisher.

1. Replication



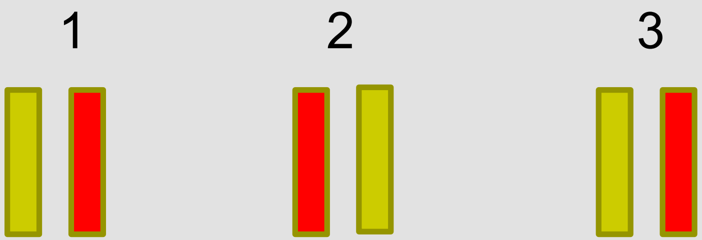
Sample size =3

2. Randomization



A “completely randomized design

3. Blocking



A Randomized block design

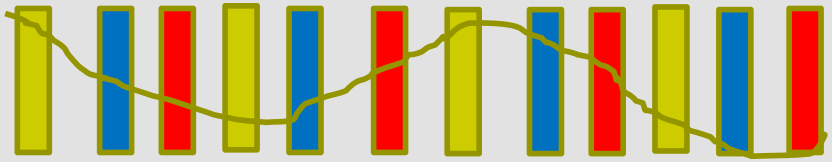
.

Basic designs: Completely randomised and randomised block experiments

First in theory, then real examples

A completely randomised design

There can be any number of treatments (3 here). “Treatment” is a *fixed effect factor*

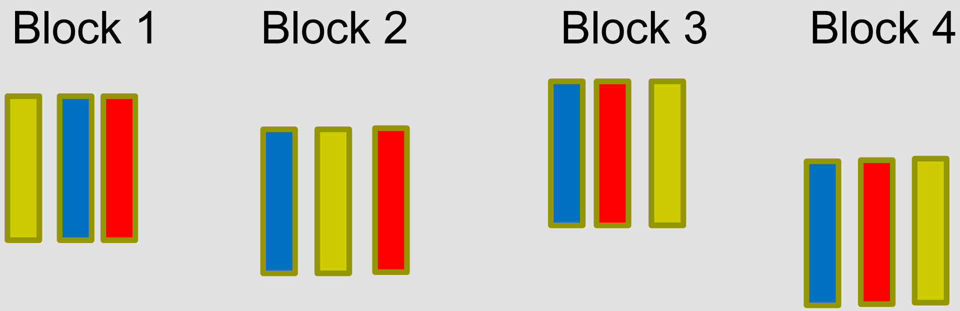


This has one fixed effect factor “treatment” (three treatments)
Statistical analysis is a one-way ANOVA

ANOVA

Source	DF	SS	MS	F	P
Treatment	2				
Error	9				
Total	11				

A randomised block design



Each block is randomised separately.
It has two factors “Treatment” (fixed effect) and “Block” (random effect).

The statistical analysis is a 2-way ANOVA without interaction.

Source	DF
Blocks	3
Treatment	2
Error	6
Total	11

Each block has a single subject on each treatment.
Blocks can be separated in space and time.
Animals within a block should be matched

The research environment

"Our lives and the lives of animals are governed by cycles,
Seasons,
reproductive cycle, weekend-working days,
cage change/room sanitation cycle,
and the diurnal rhythm.

Some of these may be attributable to routine husbandry,
the rest are cycles, which may be affected by husbandry procedures.

Other issues to be considered are
in-house transport,
Environmental effects of cage location,
The physical environment inside the cage (wet/dry),
The acoustic environment audible to animals,
The olfactory environment, materials in the cage, cage complexity, feeding regimens, kinship and interaction with humans."

Barometric pressure
Lunar cycle?
Nevalainen T. Animal husbandry and experimental design. ILAR J 2014;55(3):392-8.

The randomized block design

- More powerful (better control of the research environment)
- More convenient.
  - Work spread over time
- Less subject to bias
  - Separate randomizations for each block
  - Discourages use of historical controls or adding on of additional treatment groups post-hoc
- Makes good use of heterogeneous material
  - Animals within a block matched

Factorial designs

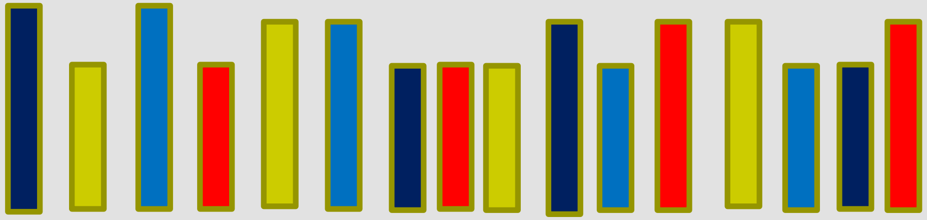
(*By using a factorial design*)”.... an experimental investigation, at the same time as it is made more comprehensive, may also be made more efficient if by more efficient we mean that more knowledge and a higher degree of precision are obtainable by the same number of observations.”

R.A. Fisher, 1960

“..we should, in designing the experiment, artificially vary conditions if we can do so without inflating the error.

Cox, DR 1958

Basic designs: Completely randomised and randomosed block 2x4 factorial experiments

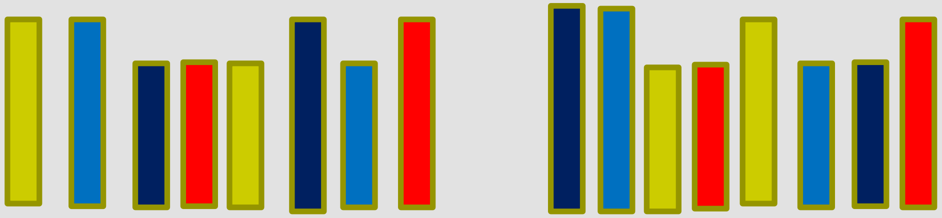


2 genders (tall/short) x 4 treatments (black, blue, brown, red)

A completely randomised "Factorial" design with four treatments and two "genders, (male tall) and females (short), all fixed effects

ANOVA 2-way with interaction

Source	DF
Treatment (T)	3
Gender (G)	1
TxG	3
Error	8
Total	15



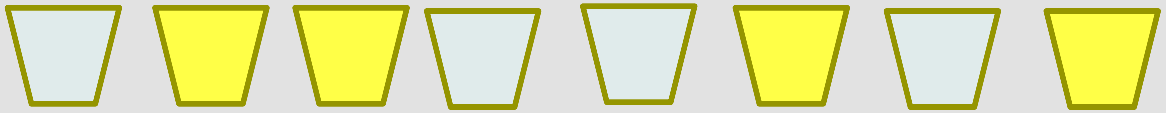
Each block has a single representative of each gender and treatment

A 4 (treatments)x2 (genders) factorial design (fixed effects) in two blocks (random effects). Analysis 3-way ANOVA with two fixed and one random factor (the blocks).

Source	DF
Blocks	1
Treatments	3
Gender	1
TxG	3
Error	7
Total	15

Randomisation, the p-value and the significance level: the basis of statistical testing (RA Fisher and the tea tasting experiment)

A lady claims that she can tell whether the milk is put in the cup before or after the tea. An experiment is set up to test this. Eight cups of tea are prepared, with four TM and four MT. They will be presented to the lady in random order and she will indicate which type they are.



Number of ways of choosing four cups out of eight cups =

$\frac{n!}{r!(n-r)!}$ = 1680/24 = 70. Only 1/70 is right, so if she does it correctly p=0.014

A 5% significance level is often chosen for making a decision to accept the results as not due to chance, but this is entirely arbitrary.

P-value. Probability of getting a result as extreme as, or more extreme than the observed one in the absence of a true effect

NHST (null hypothesis significance testing) has some critics

Recently the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values because the statistics were too often used to support lower-quality research.

Original Articles

Life After NHST: How to Describe Your Data Without "*p*-ing" Everywhere

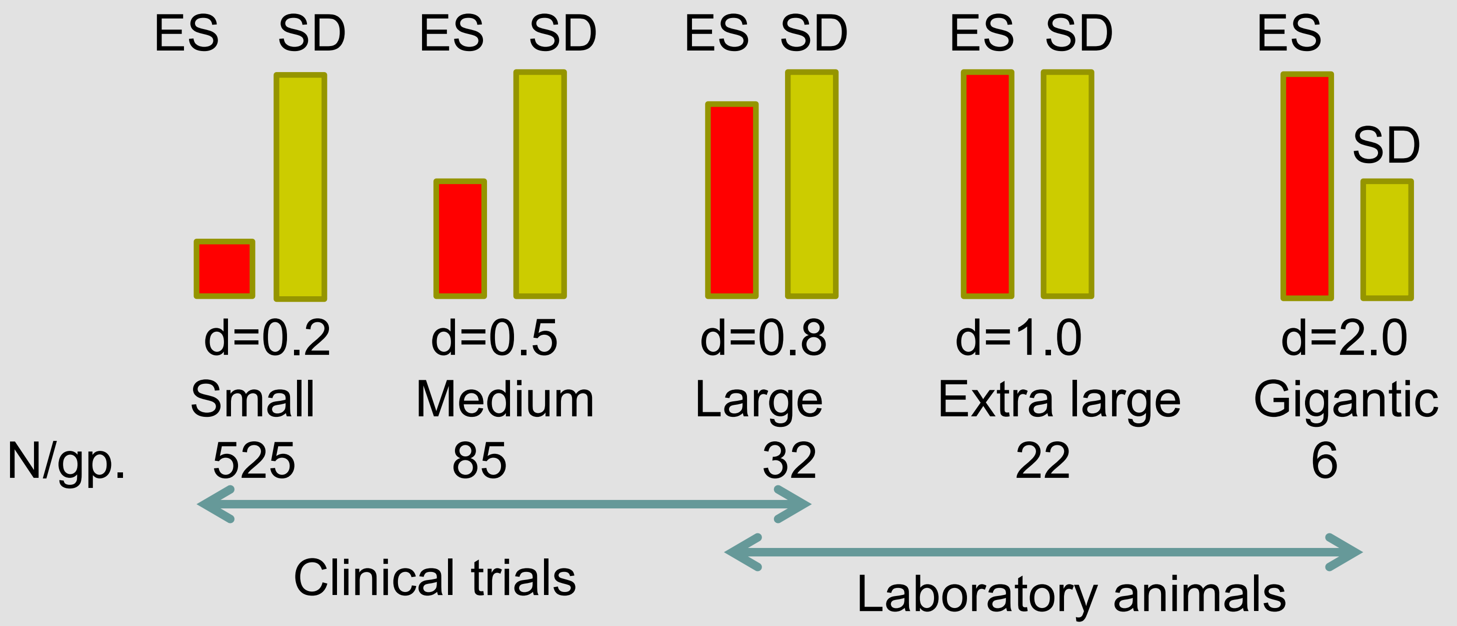
Jeffrey C. Valentine, Ariel M. Aloe & Timothy S. Lau

Pages 260-273 | Published online: 04 Aug 2015

The “standardised effect size”, SES, or Cohen’s *d*

A measure of the magnitude of a difference between means **in units of standard deviations**. A partial replacement of NHST?

d= ES/SD_p Effect size= response in standard deviation units



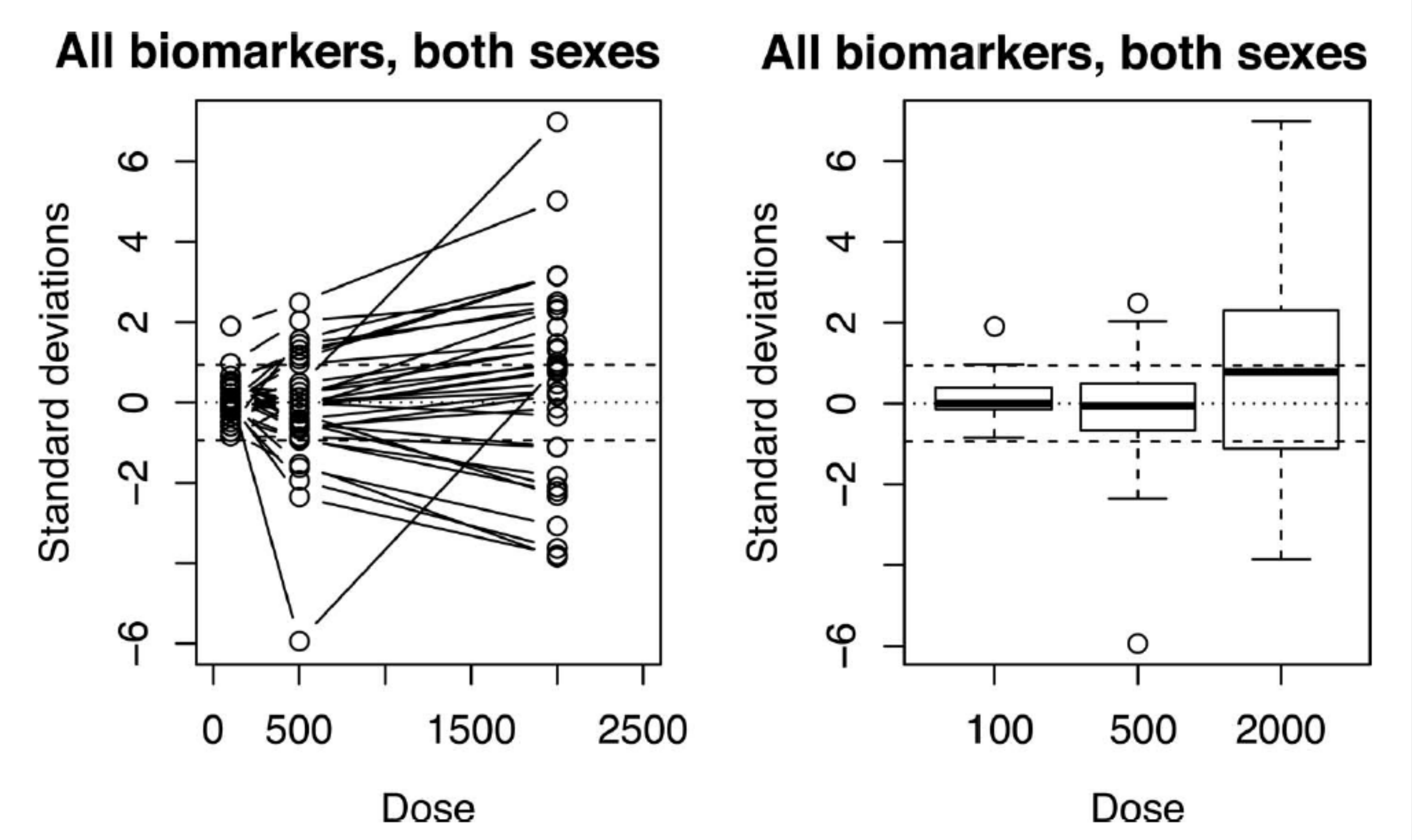
	d=0.2	d=0.5	d=0.8	d=1.0	d=2.0
	Small	Medium	Large	Extra large	Gigantic
N/gp.	525	85	32	22	6

Clinical trials

Laboratory animals

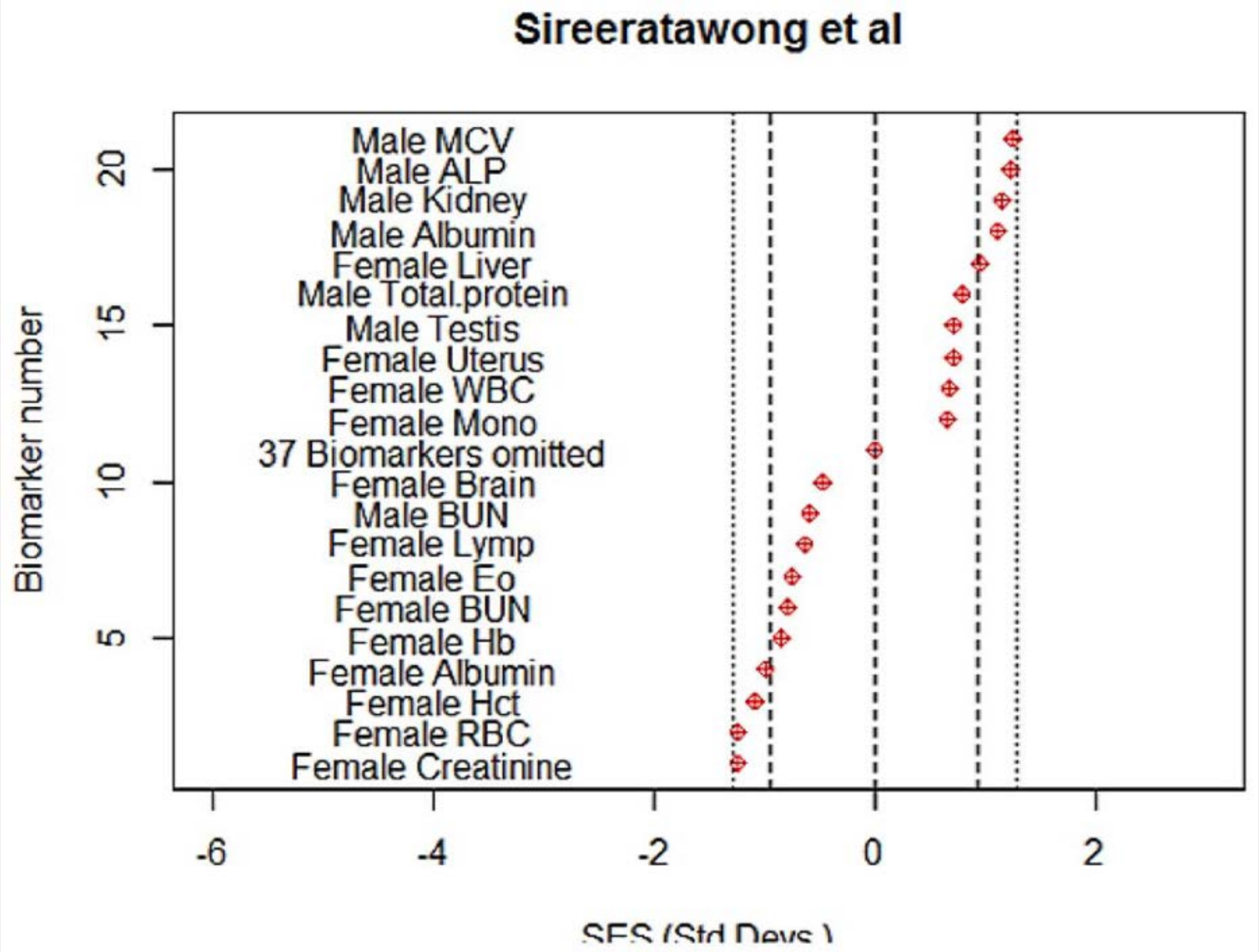
Example: Mean treated=3.30, mean control =1.55 , diff= 1.75. SD= 0.89
So *d*=1.75/0.89=1.96 SDs

Use of SESs in describing results of toxicity tests. All results converted from original units to SESs.



Michael Festing
DOI: 10.1177/0192623313517771
Toxicol Pathol published online 31 January 2014

Highlight the most changed biomarkers of toxicity



Michael Festing
DOI: 10.1177/0192623313517771
Toxicol Pathol published online 31 January 2014

Use of SES to study toxicity of GM crops in rats

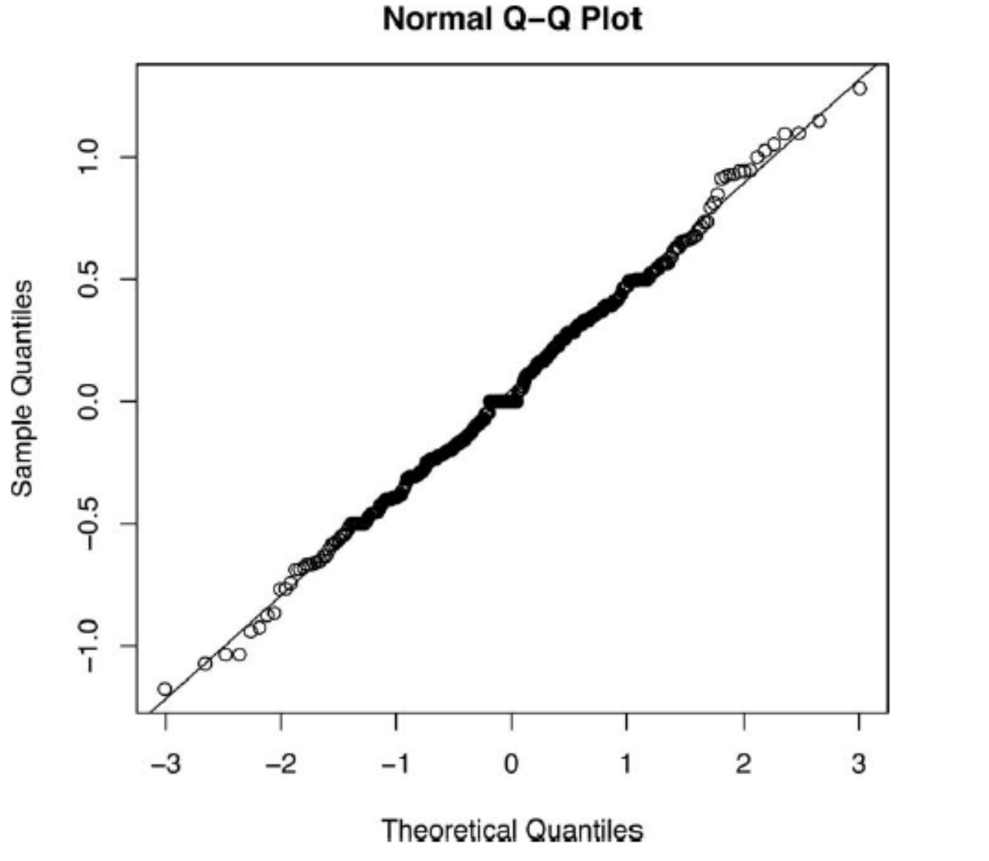


FIGURE 4. qq plot for the combined biomarkers for SES1 and SES2 (both comparing GM with non-GM) and SES5 and SES7 (comparing non-GM with non-GM), a total of 380 SESs (Mackenzie et al. 2007). A *p* value of .544 using the Shapiro Wilks test suggests that there is no evidence that the SESs deviate from a normal distribution, although

Michael Festing
DOI: 10.1177/0192623313517771
Toxicol Pathol published online 31 January 2014

This study has produced 380 differences between hematology, clinical biochemistry and organ weights in animals fed on GM corn and non GM corn. When plotted on a normal probability plot they are normally distributed. No evidence of toxicity.

Three types of experiment

- Pilot study
  - Logistics and preliminary information
- Exploratory experiment
  - Aim is to provide data to generate hypotheses
  - May “work” or “not work”
  - Often many outcomes
  - Statistical analysis may be problematical (many characters measured, data snooping). p-values may not be correct
  - “The Texas sharp-shooter problem”
- Confirmatory experiment (Gold standard)
  - Formal hypothesis stated *a priori.* Randomised controlled experiment.
  - Various designs including “completely randomised” and “randomised block” designs.

A well designed confirmatory experiment

- Clearly stated objectives
- Absence of bias
  - Experimental unit, randomisation, blinding
- High power
  - Low noise (uniform material, blocking, covariance)
  - High signal (sensitive subjects, high dose)
  - Large sample size
- Wide range of applicability
  - Replicate over other factors (e.g. sex, strain): factorial designs
- Simplicity
- Amenable to a statistical analysis
  - Planned with the design

Internal validity

External validity

Real Example 1.
A completely randomised (CR) design

Purpose of the study:

Do MCA and Urethane increase micronuclei in the peripheral blood of BALB/c female mice.

12 mice per group.
Treatments were assigned to mice at random.
Micronuclei were counted blind using the laser scanning cytometer.

Problems with a CR design:

1. May not be possible to obtain large numbers of animals of uniform weight, age etc.
2. May not be able to house them them in a uniform environment
3. May not be able to measure them in a uniform environment

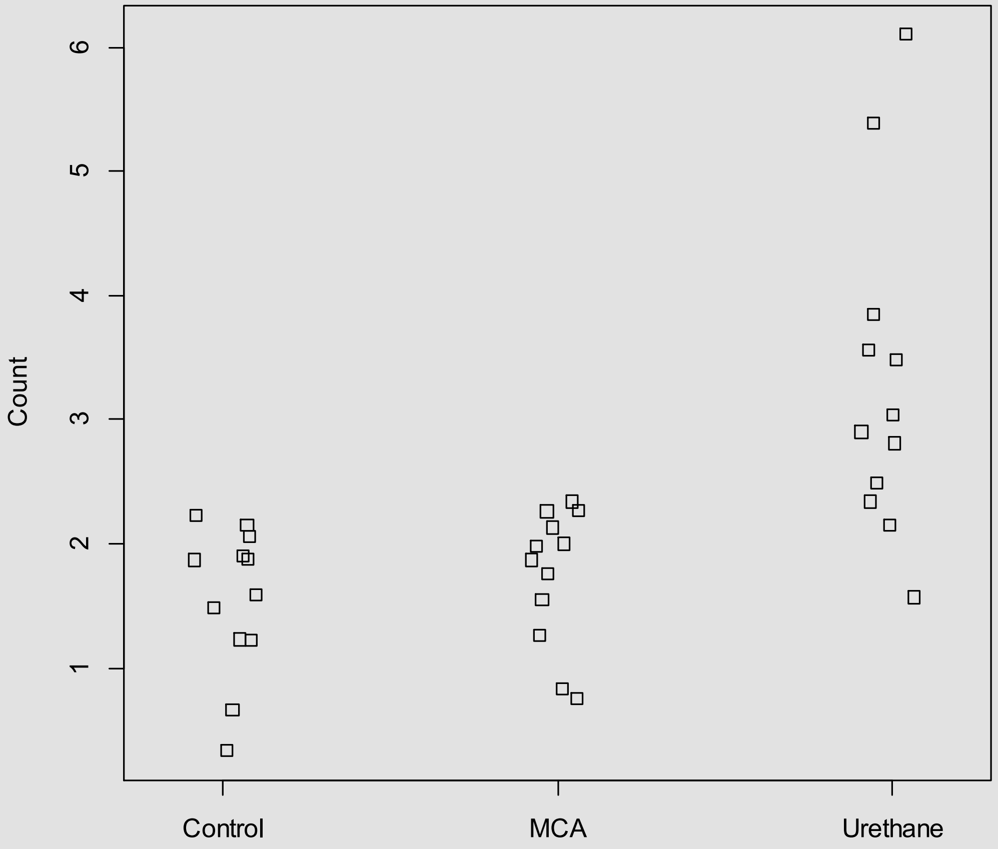
So, inter-individual variability may be increased, and power decreased, because: SD increased.

However, the design is simple and is widely used.

Animal	Treatment	Count	
1	Urethane	3.48	
2	Control	1.9	
3	Control	1.23	
4	MCA	1.26	
5	MCA	2.34	
6	Urethane	5.39	*
7	Control	2.06	
8	Urethane	2.34	
9	MCA	1.55	
10	MCA	2.26	
11	Control	1.87	
12	Control	0.66	
13	Urethane	3.85	
14	Urethane	1.57	
15	MCA	2.00	
16	Control	2.15	
17	MCA	2.13	
18	MCA	2.27	
19	Urethane	3.56	
20	MCA	1.98	
21	MCA	1.76	
22	Control	1.22	
23	Urethane	6.10	*
24	Control	1.59	
25	Control	1.88	
26	Control	2.23	
27	MCA	1.87	
28	Control	0.33	
29	Urethane	2.15	
30	MCA	0.83	
31	Urethane	2.81	
32	Control	1.48	
33	Urethane	2.9	
34	MCA	0.75	
35	Urethane	2.49	
36	Urethane	3.04	

Statistical analysis

Plot individual points



“jitter” added so points separated horizontally

ANOVA assumptions:

1. Equal variances
2. Residuals have normal distribution
3. Independent experimental units.

What about the two outliers?
(do they make a difference to the conclusions?)

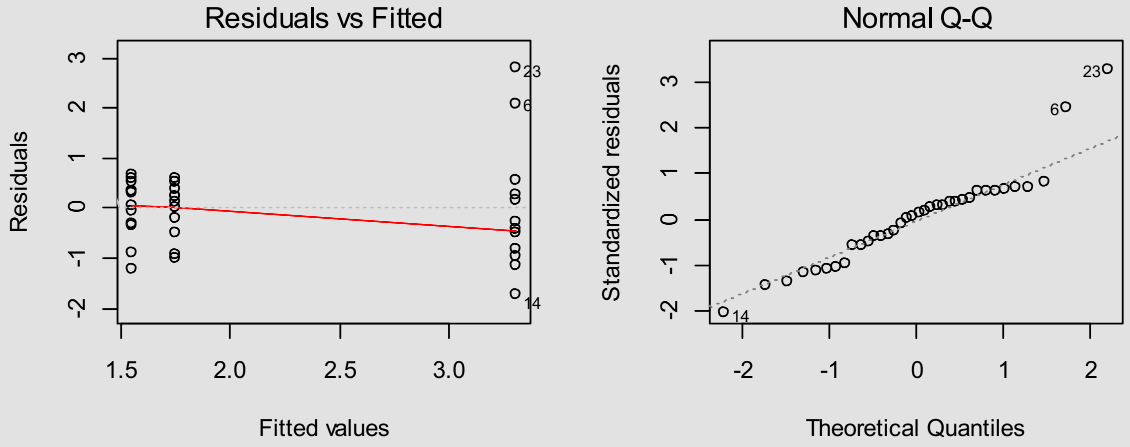
A trial ANOVA (to look at residuals)

```
Source          Df       SS          MS         F         P
Treatment       2       22.196     11.0982     13.997 <0.001
Residuals       33      26.165      0.7929
Total           35      48.361

Pooled sd= sqrt(.7929) = 0.890
```

Pooled variance

Residuals diagnostic plots

aov(Count ~ Treatment)



Assumptions for a parametric analysis:

1. Normal distribution of residuals
2. Homogeneous variances
3. Observations are independent (part of the design)

Should be a scattering of points with no pattern

Points should fall on a straight line

Means and standard deviations

Treat.	mean	sd	n	Post-hoc comparisons*
Control	1.55	0.596	12	a
MCA	1.75	0.546	12	a
Urethane	3.30	1.313	12	b

Pooled sd = 0.89 (from sqrt EMS in ANOVA)

*post-hoc comparisons done using Tukey's test

Standardised effect sizes/Cohen's *d*:

***d (SES)*= (Diff. between means)/pooled SD)**

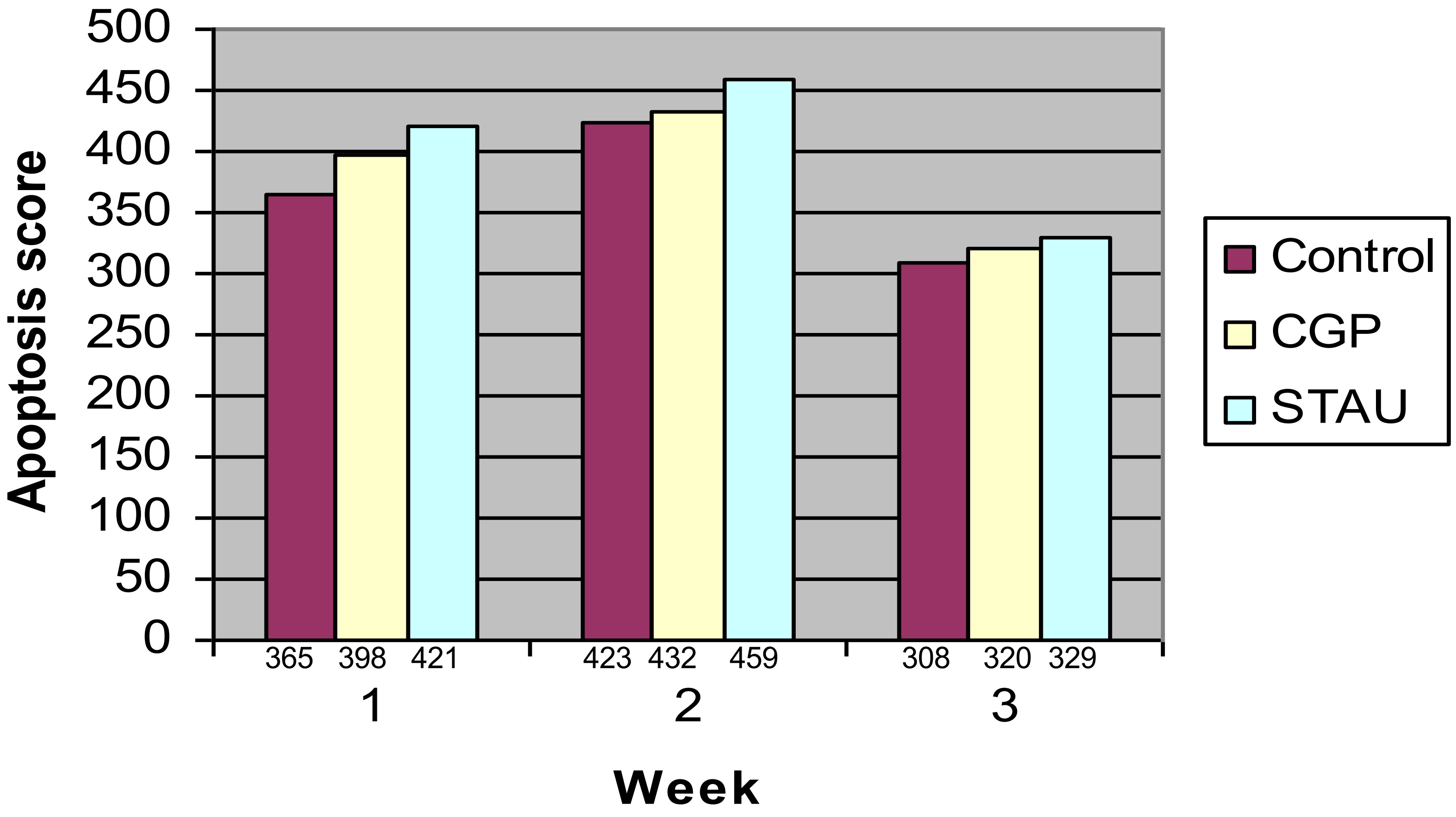
SES: MCA = (1.75-1.55)/0.89 = 0.22

Urethane= (3.30-1.55)/0.89= 1.96

Note: I have been inconsistent & used SES and Cohen's *d* for the same thing

Example 2. A randomised block experiment

Do “CPG and STAU increase apoptosis in rat thymocytes?
Experimental unit is a dish of thymocytes



Advantages of randomised block designs

- If blocked in time, provides some assurance of repeatability
  - In-vitro experiments often say “We repeated the experiment three times”
- More powerful than CR design. Better control of variation.
  Two animals treated at same time and housed in adjacent cages likely to be more similar than two treated at different times and housed on different shelves.
- More convenient: can be done a bit at a time
- Less susceptible to faulty randomisation

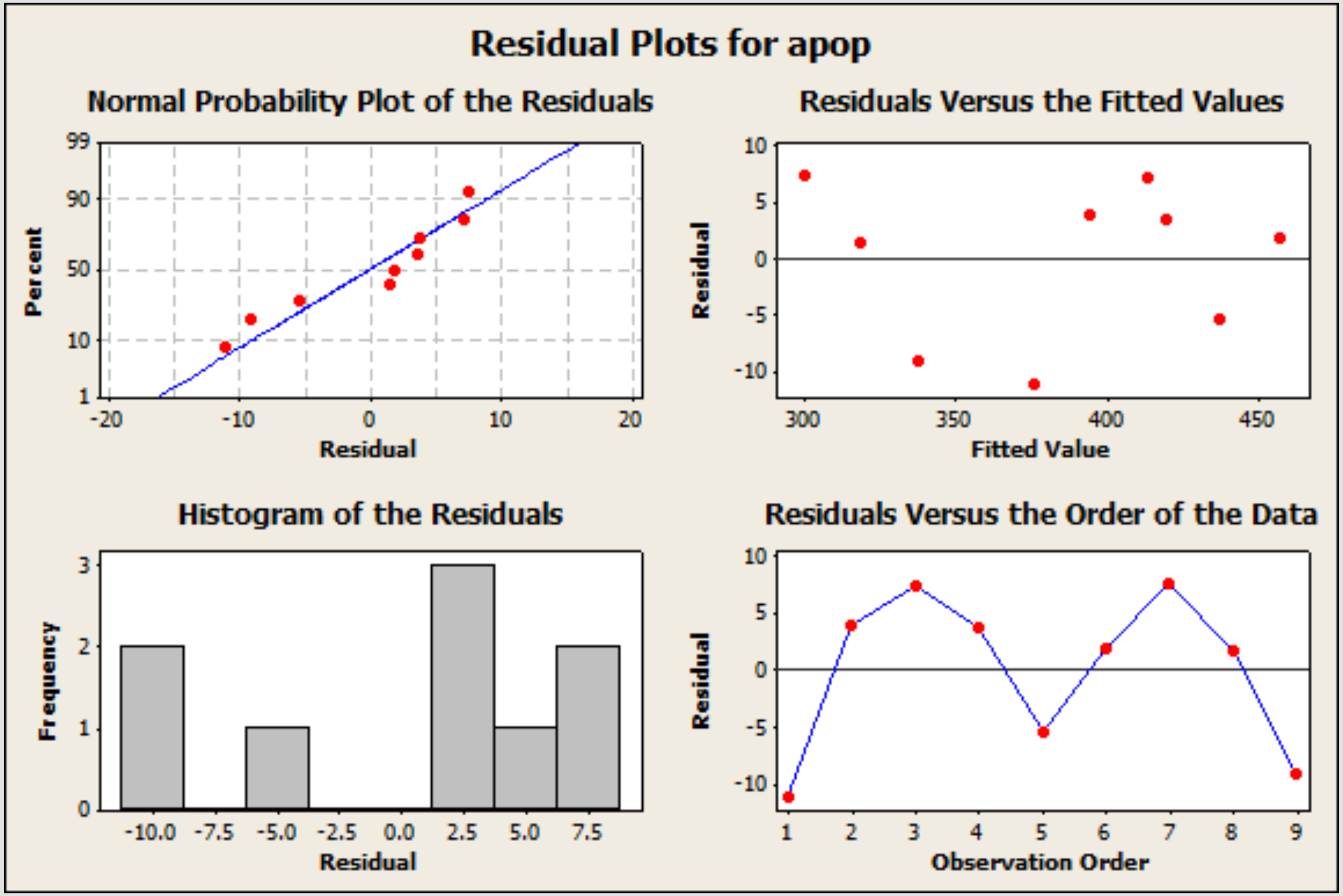
- Disadvantages:
- Not so good with several missing observations /unequal sample sizes (a few tolerated)
- Requires a 2-way ANOVA without interaction

ANOVA (MINITAB)

```
Week     random         3  1, 2, 3
Drug     fixed          3  C, CGP, STAU


Analysis of Variance for apop

Source   DF       SS       MS       F      P
Week      2  21764.2  10882.1  114.82  0.000
Drug      2   2129.6   1064.8   11.23  0.023
Error     4    379.1     94.8
Total     8  24272.9


S = 9.73539   R-Sq = 98.44%
```

An estimate of the pooled variance

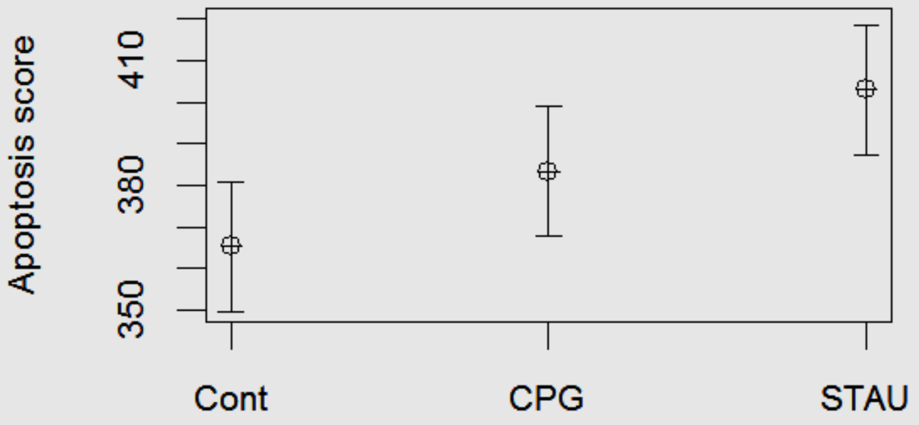
Residuals plots (done with MINITAB)



Means etc

Group	Mean
C	365
CPG	383
STAU	403*

Pooled SD= 9.7



Post-hoc comparison:
Dunnett Simultaneous Tests
Response Variable apop
Comparisons with Control Level
treat = C subtracted from:

```
      Difference      SE of        Adjusted
treat    of Means  Difference  T-Value    P-Value
CPG         18.00       7.949   2.264    0.1419
STAU        37.67       7.949   4.739    0.0155
```

Standardised effect sizes
CPG = (383-365)/9.7 = 1.85
STAU = (403-365)/9.7 = 3.91

Example 3
Factorial designs

(*By using a factorial design*)”.... an experimental investigation, at the same time as it is made more comprehensive, may also be made more efficient if by more efficient we mean that more knowledge and a higher degree of precision are obtainable by the same number of observations.”

R.A. Fisher, 1960

“..we should, in designing the experiment, artificially vary conditions if we can do so without inflating the error.

Cox, DR 1958

Example 3.Factorial designs are widely used but often incorrectly analysed

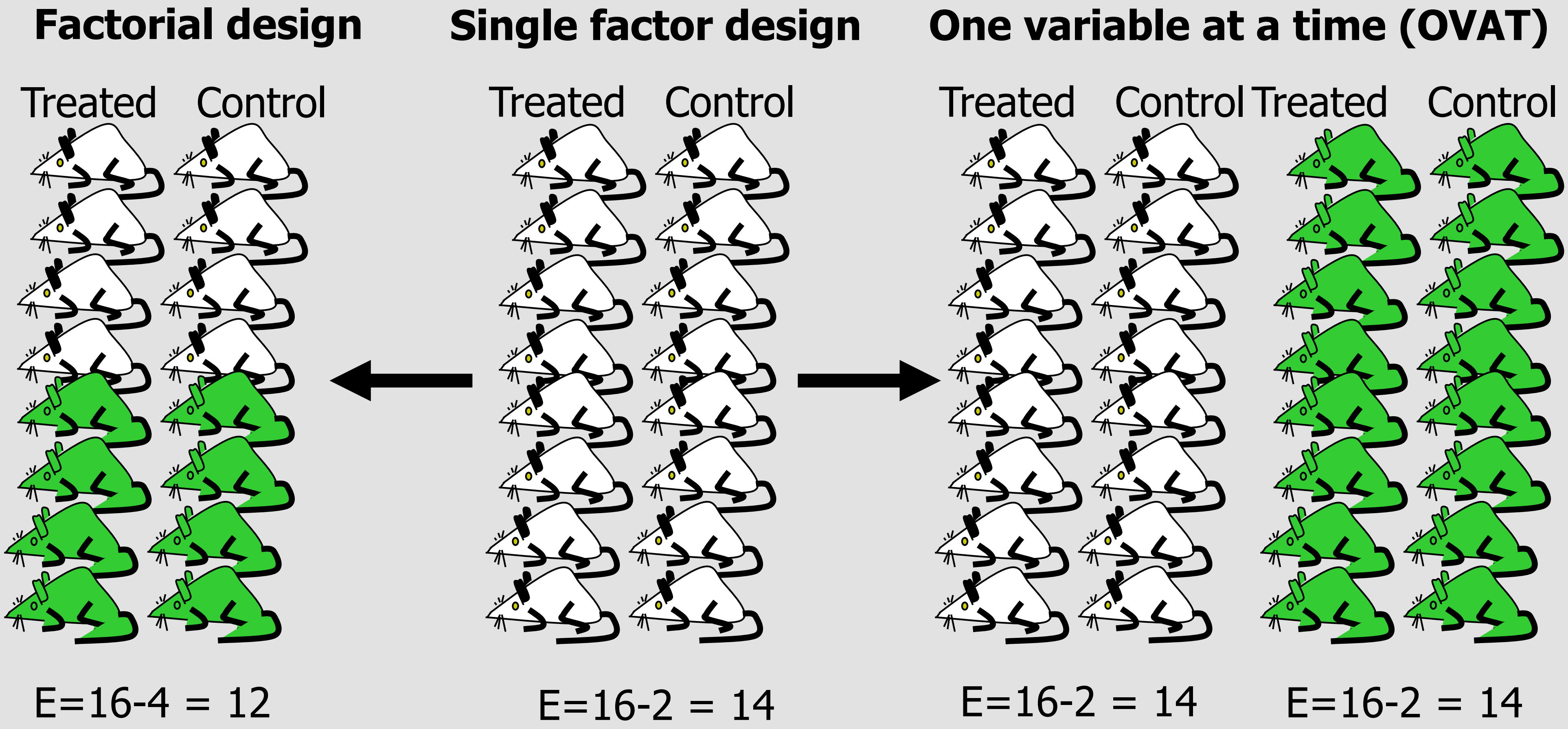
Number of studies	513 (Neuroscience papers)
Factorial designs	153 (30%)
Correctly analysed	78 (50%)

Niewenhuis et al (2011) Nature Neurosci. 14:1105

Need a 2-way ANOVA **with interaction**

Example 3. Factorial "designs"

(they are really an arrangement of treatments)



Factorial design

Treated Control

E=16-4 = 12

Single factor design

Treated Control

E=16-2 = 14

One variable at a time (OVAT)

Treated Control Treated Control

E=16-2 = 14 E=16-2 = 14

Example 3a. Effect of chloramphenicol on RBC counts (2000μg/kg)

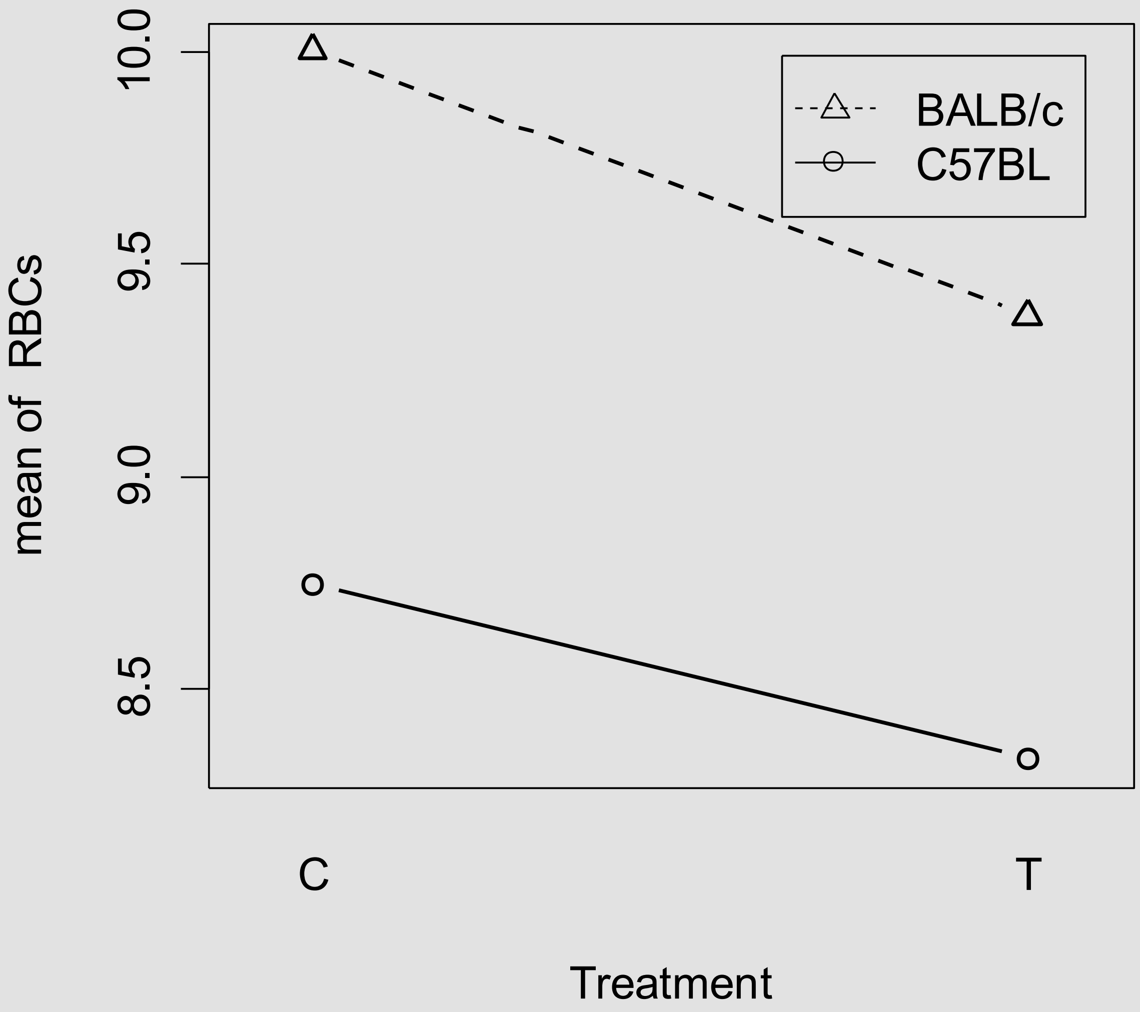
No interaction

Want to know:

1. Does treatment have an effect on RBC counts
2. Do strains differ in RBC counts
3. Do strains differ in their response (interaction)

Strain	Control	Treated	Strain means
BALB/c	10.10	8.95	
	10.08	8.45	
	9.73	8.68	
	10.09	8.89	9.37
C57BL	9.60	8.82	
	9.56	8.24	
	9.14	8.18	
	9.20	8.10	8.86
Treat. Mean	9.69	8.54	

Example 3a. No interaction



Example 3a. No interaction

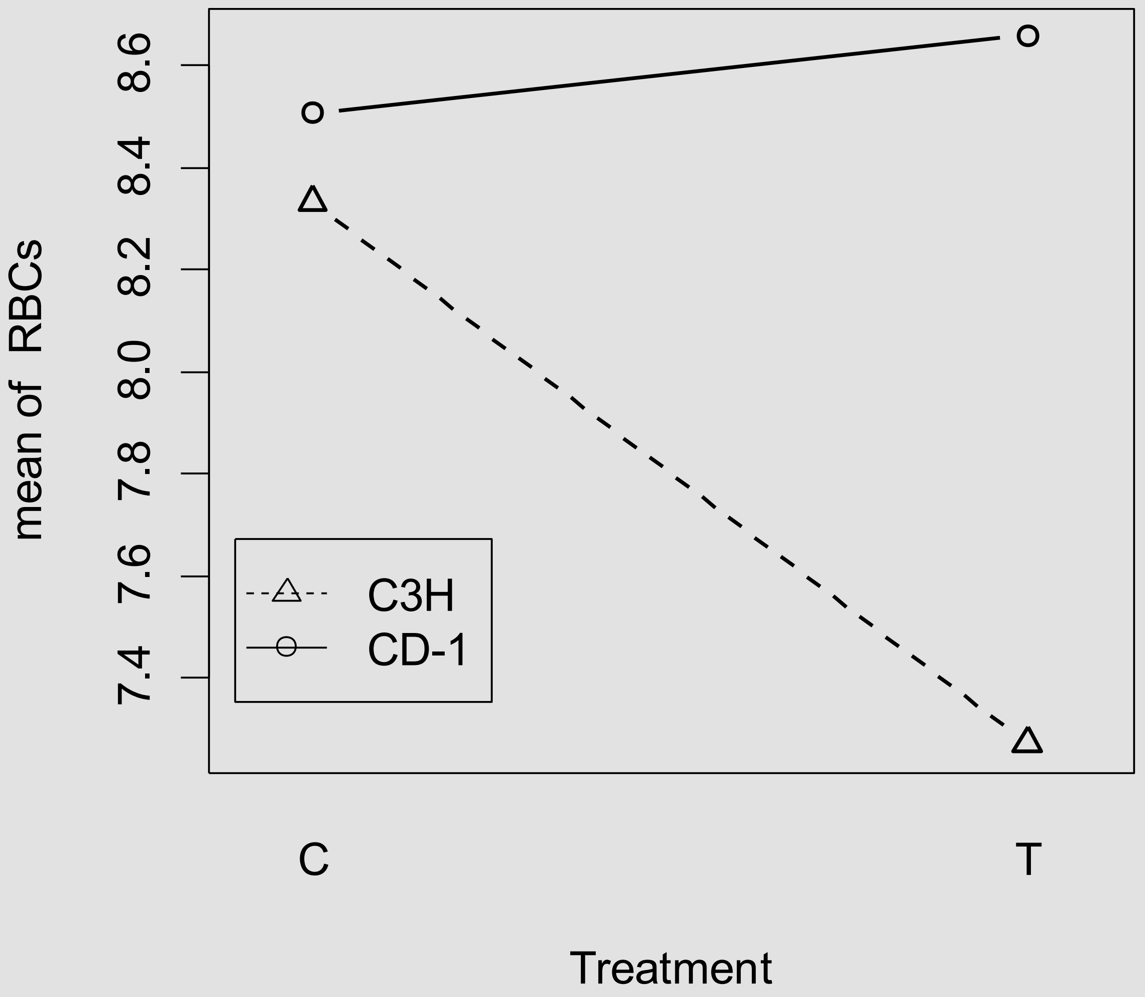
```
Analysis of Variance Table

Response: RBCs
                 Df Sum Sq Mean Sq F value    Pr(>F)    
Treatment         1 1.0661  1.0661 17.1512  0.001367 ** 
Strain            1 5.2785  5.2785 84.9232 8.595e-07 ***
Treatment:Strain  1 0.0473  0.0473  0.7611  0.400108    
Residuals        12 0.7459  0.0622                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 
' ' 1 
> 
```

Example 3b. Effect of chloramphenicol (2000mg/kg) on RBC count

Significan Interaction

Strain	Control	Treated	Strain means
C3H	7.85	7.81	
	8.77	7.21	
	8.48	6.96	
	8.22	7.10	7.80
CD-1	9.01	9.18	
	7.76	8.31	
	8.42	8.47	
	8.83	8.67	8.58
Treatment means	8.42	7.96	

Example 3b. Interaction



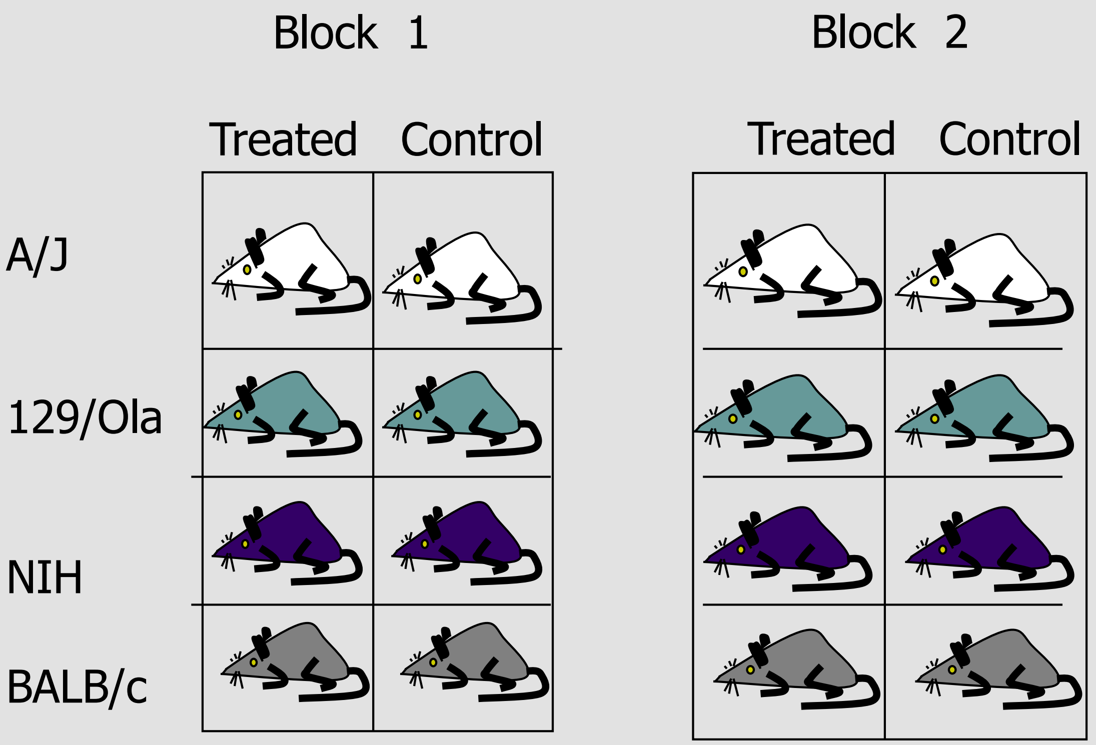
Example 3b
ANOVA with significant interaction

```
Analysis of Variance Table

Response: RBCs
                 Df  Sum Sq Mean Sq F value   Pr(>F)   
Strain            1 0.82356 0.82356  4.4302 0.057057 . 
Treatment         1 2.44141 2.44141 13.1330 0.003489 **
Strain:Treatment  1 1.47016 1.47016  7.9084 0.015686 * 
Residuals        12 2.23077 0.18590                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1
>
```

Example 4. A 2x4 factorial design in two blocks.

Effect of diallyl sulphide (DS) on the activity of liver Gst in mice of four inbred strains

DS Administered by gavage in three daily doses of 0.2mg/g. to eight week old female mice



One female mouse per cage. The two blocks were separated by approximately 2 months

Example 4. A 2x4 factorial design in two blocks. Raw data

Table 1. Gst levels* from a RB experiment in two blocks separated by approximately three months.

Strain	Treatment	Block1	Block2
NIH	C	444	764
NIH	T	614	831
BALB/c	C	423	586
BALB/c	T	625	782
A/J	C	408	609
A/J	T	856	1002
129/Ola	C	447	606
129/Ola	T	719	766

* nmol conjugate formed per minute per mg of protein

Example 4.
Analysis of the results

ANOV Gst activity Score

```
Source             Df Sum Sq Mean Sq F value     Pr(>F)
Block               1 124256  124256 42.0175 0.0003398
Strain              3  28613    9538  3.2252 0.0914353 .
Treatment           1 227529  227529 76.9394 5.041e-05 *
Strain:Treatment    3  49590   16530  5.5897 0.0283197 *
Residuals           7  20701    2957
```

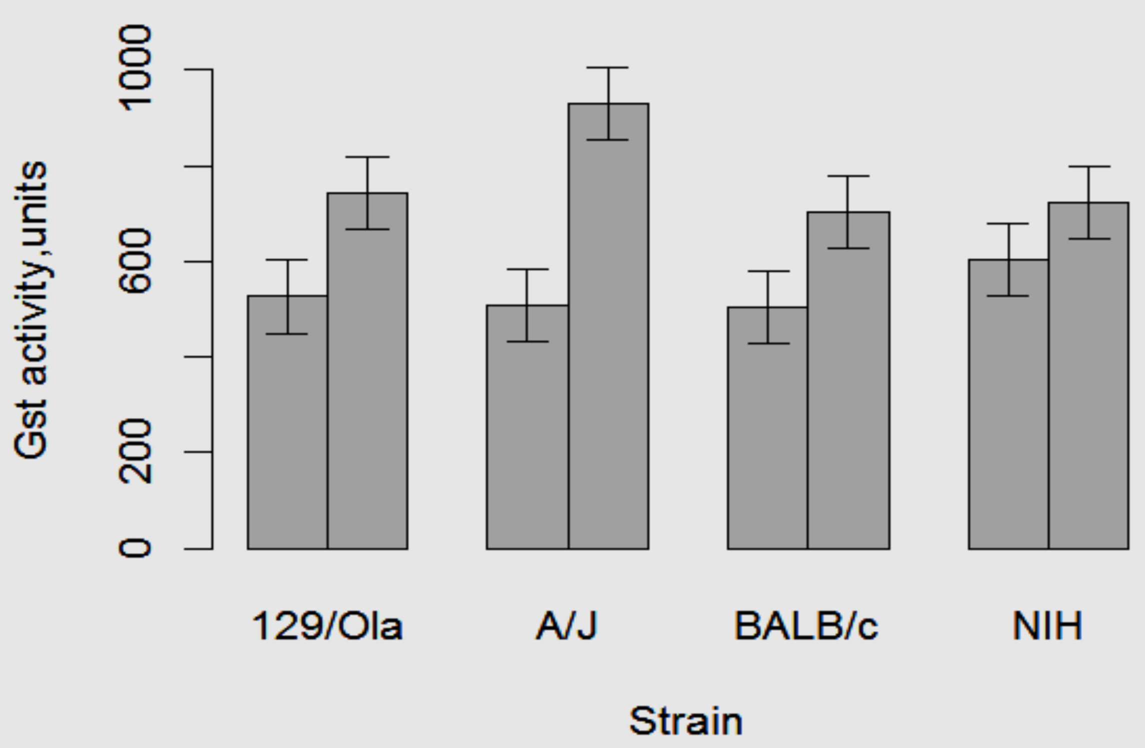
```
Treatment means
mean   data:n
C 535   8
T 774   8
```

Pooled SD = sqrt(2957) = 54.3

```
Strain means
Strain  mean   n
129/Ola 634    4
A/J     718    4
BALB/c  604    4
NIH     663    4
Pooled SD 54.3
```

SES(treatment)= (774-535)/54.3=4.40

Example 4. Mean responses in control and Diallyl Sulphide-treated animals



Error bars are least significant differences. If they overlap there is no significant difference ($p>0.05$), if they do not, then there is a significant difference ($p<0.05$)

A well designed experiment.

(Will have a formal design)

- Clearly stated objectives
- Absence of bias
  - **Experimental unit, randomisation, blinding**
- High power
  - Low noise (uniform material, blocking, covariance)
  - High signal (sensitive subjects, high dose)
  - Large sample size

Internal validity

- Wide range of applicability
  - Replicate over other factors (e.g. sex, strain): factorial designs

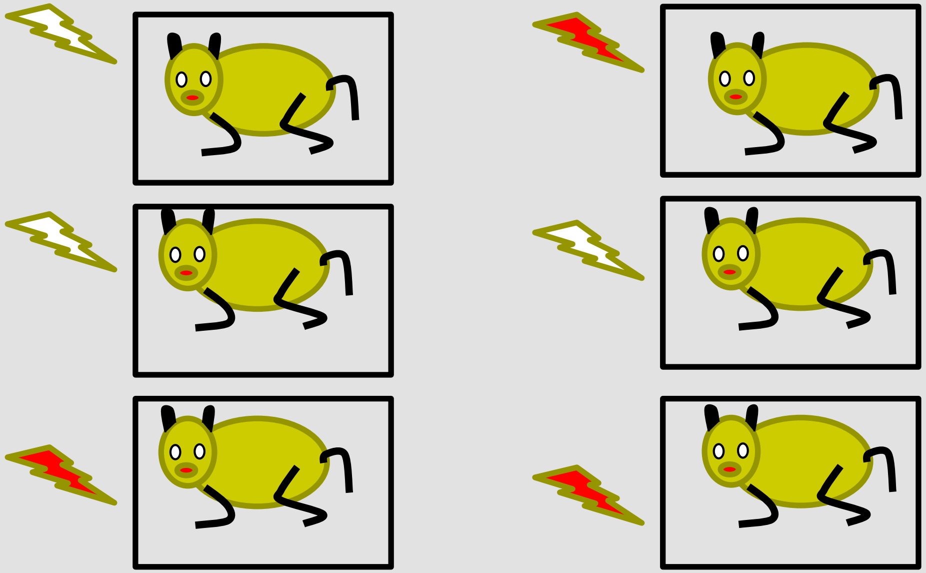
External validity

- Simplicity
- Amenable to a statistical analysis

Experimental units (EUs)

A completely randomised design

Treatments assigned to individuals at random.

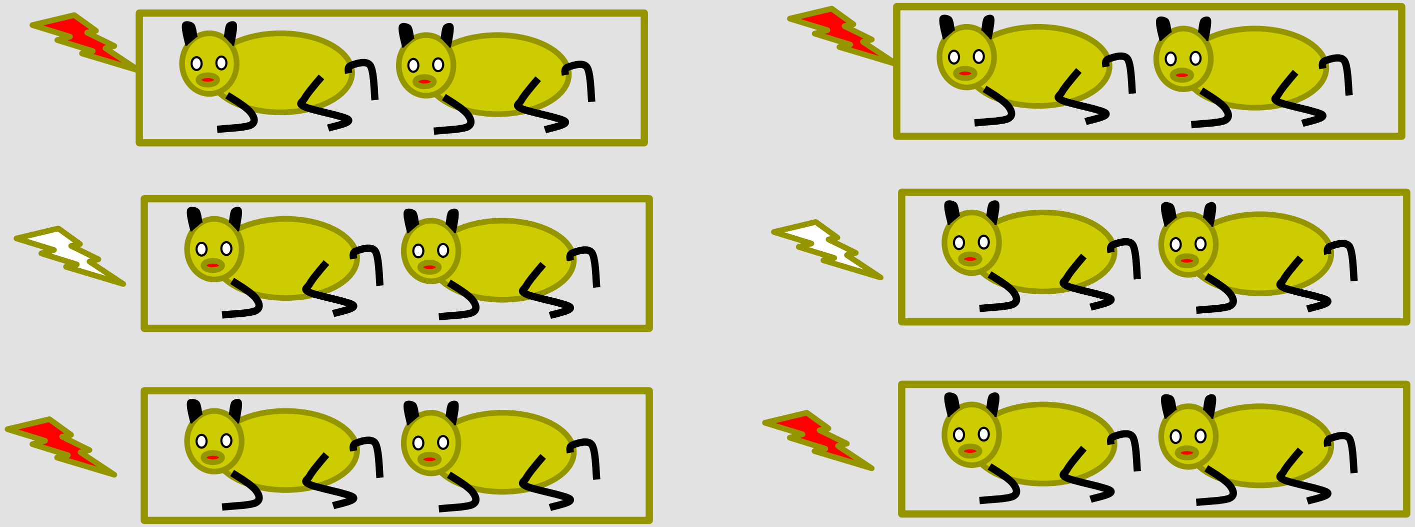


N=6

EU: Smallest division of the experimental material such that any two EUs can receive different treatments

Experimental units (EUs)

Animals within cage/pen have same treatment. A completely randomised design



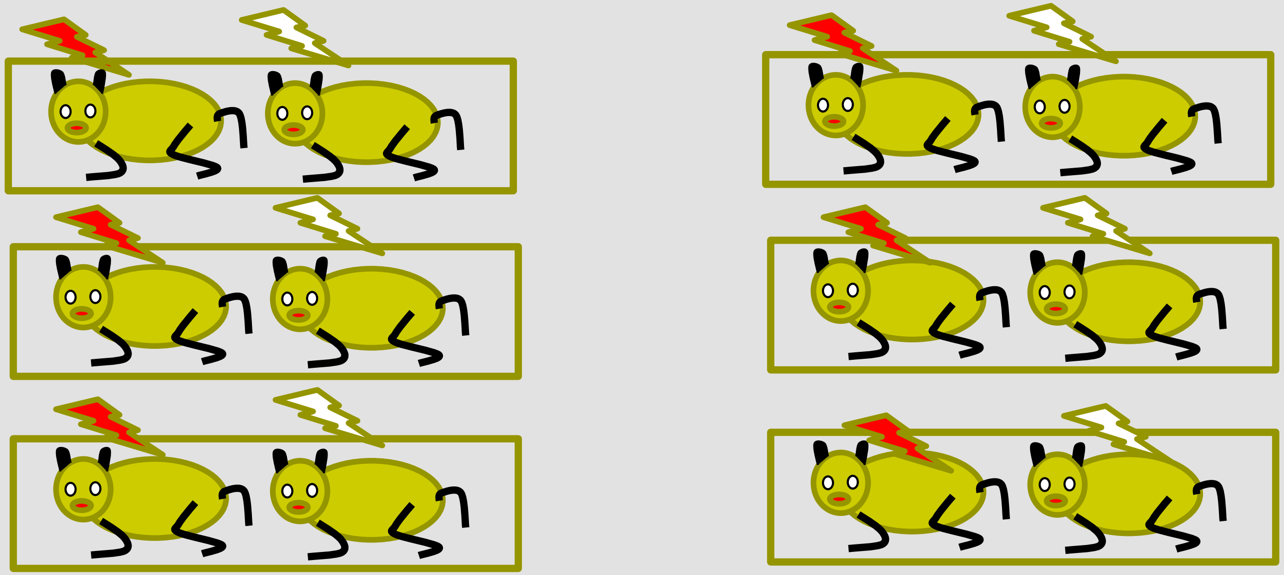
N=6

EU: A cage with two animals.

Experimental units (EU)

A randomised block design

Animal within pen have different treatments.



N=12

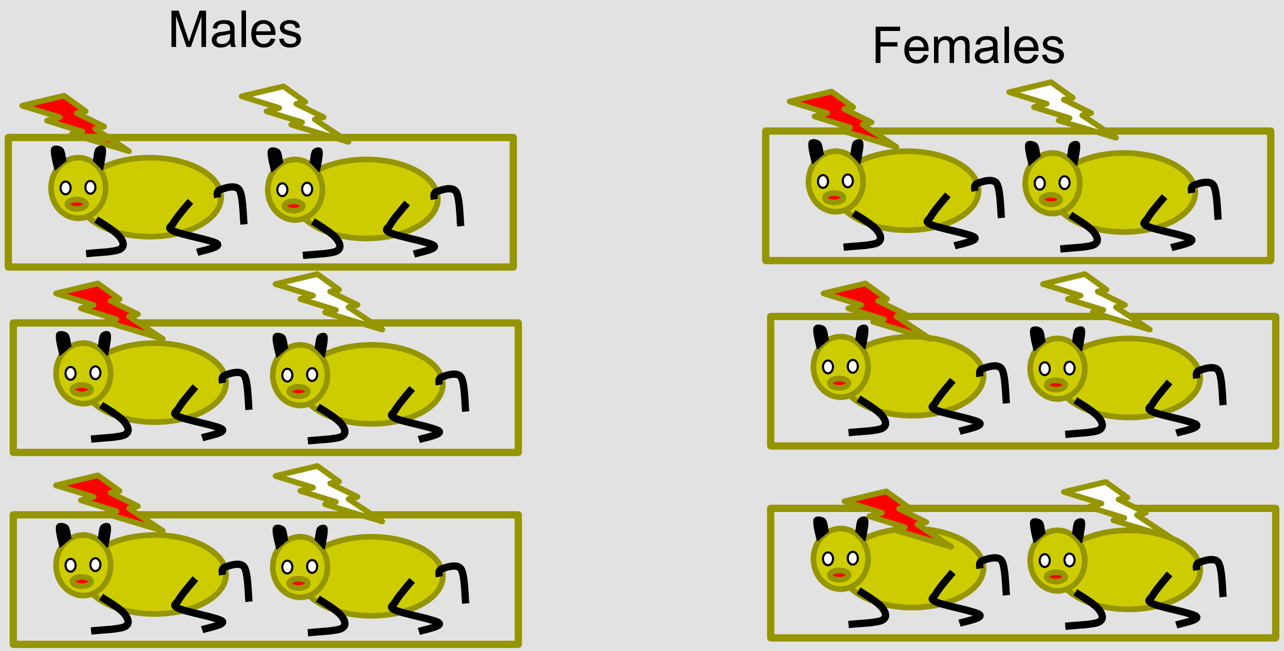
EU: Smallest division of the experimental material such that any two EUs can receive different treatments

Experimental units (EU)

A split plot design. What are the experimental units?

Animals within pen have different treatments.

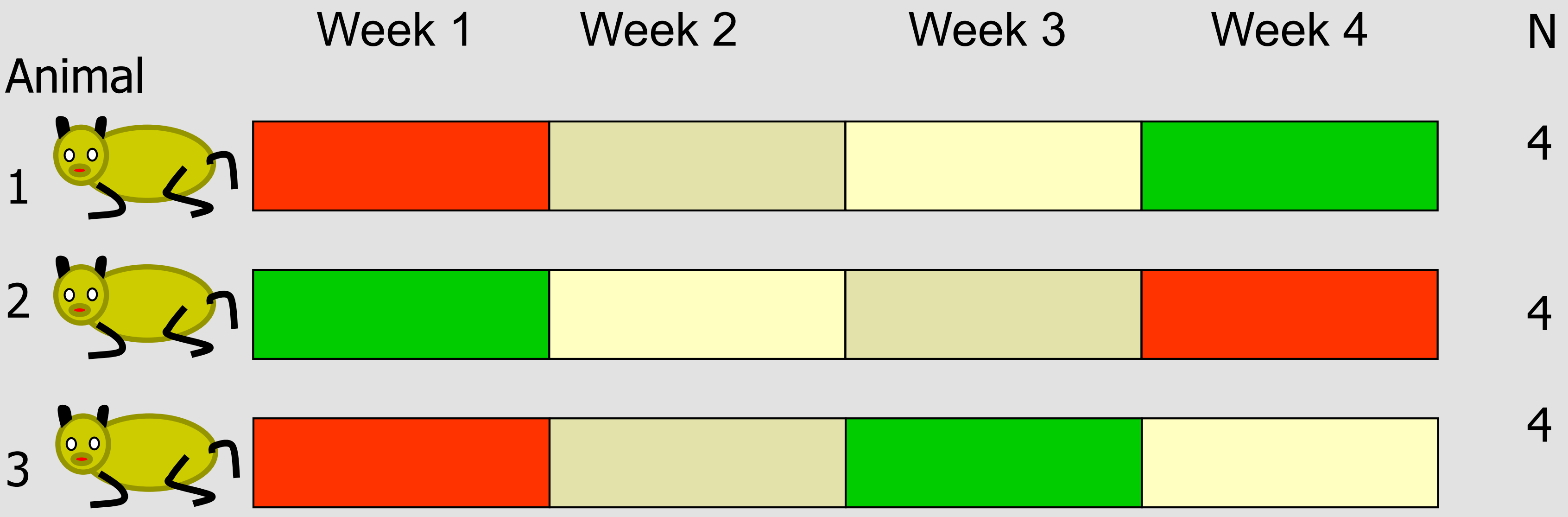
For a split-plot analysis consult a statistician



EU: Smallest division of the experimental material such that any two EUs can receive different treatments

A “Crossover” (Randomised block) design

(some authors also call this a repeated measures design)

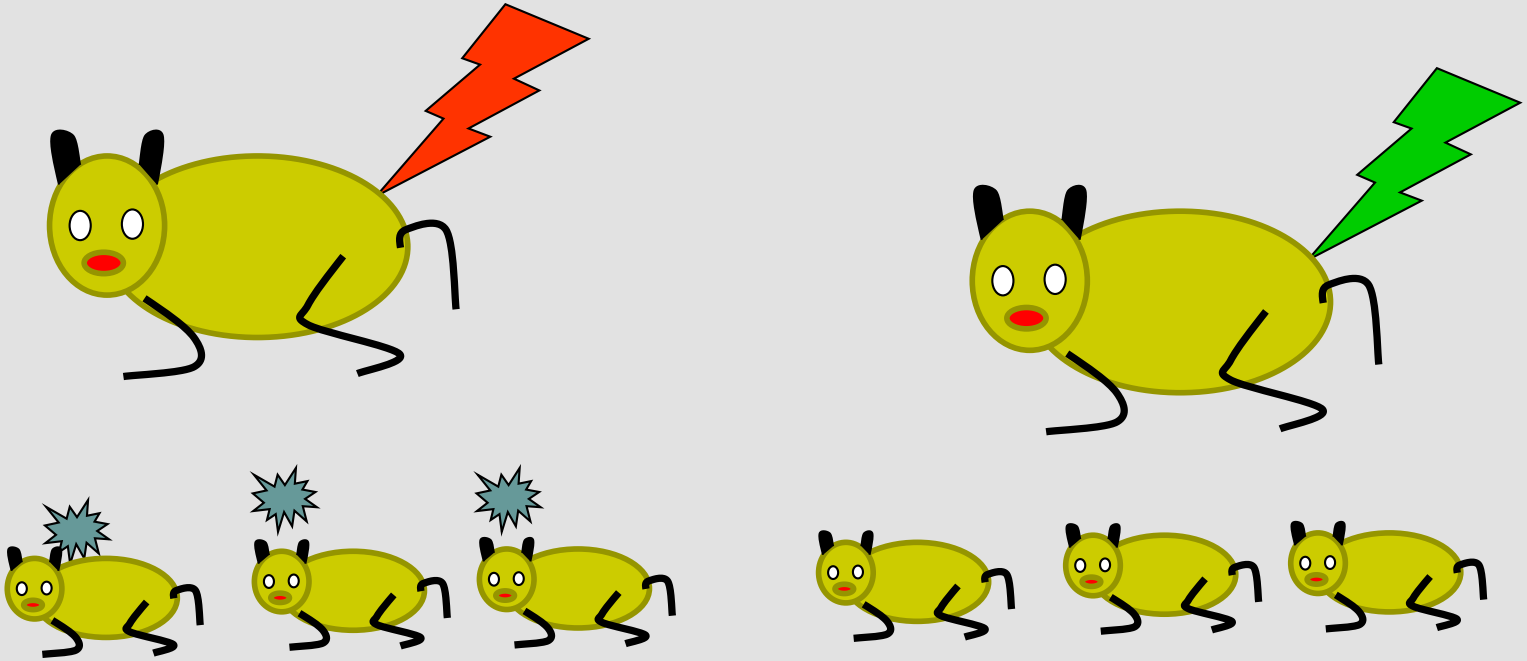


Animal	Week 1	Week 2	Week 3	Week 4	N
1	🟥	🟫	🟨	🟩	4
2	🟩	🟨	🟫	🟥	4
3	🟥	🟫	🟩	🟨	4

EU: an animal for a period of time:

N=12

Teratology: mother treated, young measured



N=2

Mother is the experimental unit.

EU learning outcome 4.
Identify the experimental unit and recognise issues of non-independence (pseudo- replication).

What is the experimental unit

An investigator wants to see whether outbred stocks are more variable than inbred strains in a test involving insect antigens.

He bought 16 BALB/c mice and compare them with 16 ICR mice looking at within-group variation in 10 different immunological tests.

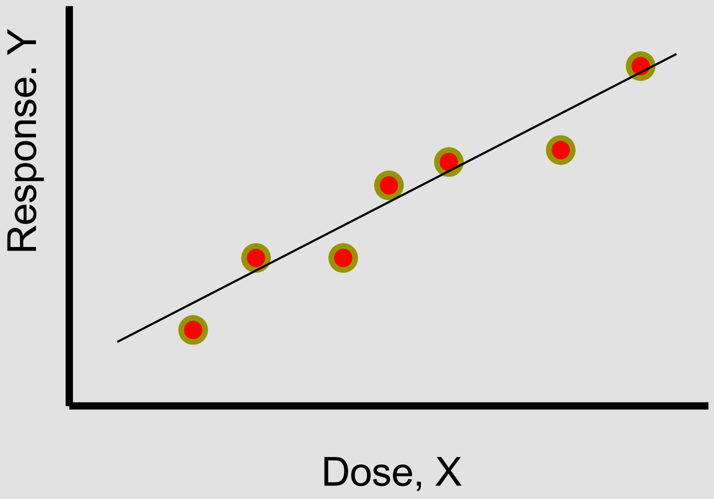
He found no difference in variability between the two groups.

He concluded that investigators could save a lot of money by using outbred stocks rather than inbred strains

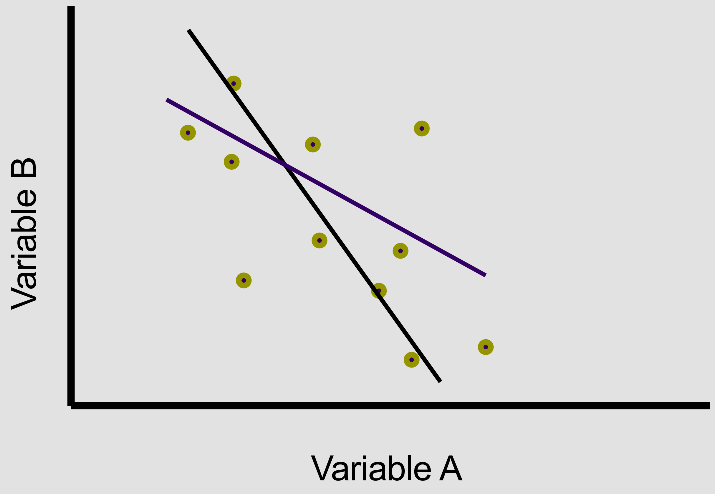
What is the experimental unit?
Other comments?

Experimental unit is the strain and there is only one of each.
Need large sample sizes to test whether two groups differ in variability

Regression and correlation



Prediction of Y from X



Association between Variables A and B

Statistical analysis should fit the purpose of the study

A Completely Randomised Design

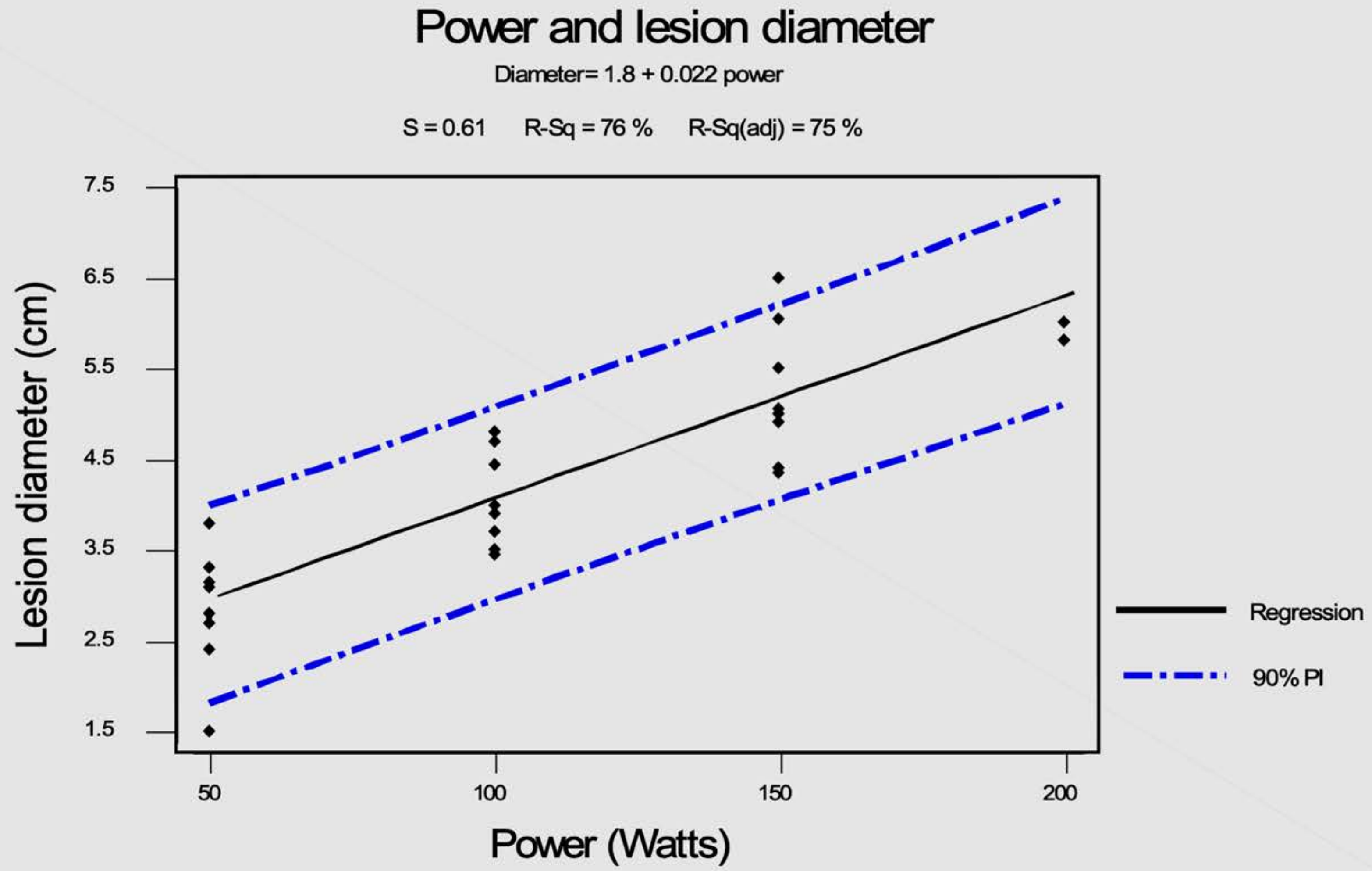
Experimental unit??

Lesion diameter following microwave treatment of liver of pigs.

Power (watts)										Mean
50	3.3	3.2	2.8	2.8	2.4	2.7	3.2	3.8	1.5	2.9
100	4.7	4.0	3.5	4.4	3.9	4.8	4.4	3.7	4.0	4.2
150	5.5	5.0	4.4	4.5	6.0	6.5	5.0	5.0		5.3
200	5.8	6.0								5.9

Lesion diameter clearly increases with power, but aim is to quantify this

Regression analysis



Randomisation

The animals are remarkably uniform. Why do we need to randomise them?

Why not assign alternatively to the two groups?

If we did this, what would be the experimental unit?

Treated group cages
1
2
3
4
5
6

Control group cages
7
8
9
10
11
12

Randomisation and blinding using EXCEL

Animal

1

2

3

4

5

6

7

8

9

10

11

12

Randomising a randomised block design

3 treatments, A, B, C.
4 blocks 1-4

Original unsorted

Treatment	Block	RandomNo
A	1	0.208
A	2	0.642
A	3	0.322
A	4	0.098
B	1	0.974
B	2	0.687
B	3	0.113
B	4	0.827
C	1	0.405
C	2	0.543
C	3	0.147
C	4	0.292

Sorted on rand() .

Treatment	Block	RandomNo
A	3	0.779
B	1	0.333
A	1	0.544
C	2	0.797
B	2	0.162
B	4	0.907
C	4	0.471
C	1	0.162
A	4	0.906
A	2	0.701
B	3	0.416
C	3	0.719

2nd. Sort on block

Treatment	Block	RandomNo	ID
			1
			2
			3
			4
			5
			6
			7
			8
			9
			10
			11
			12

Each block blinded once treatments have been given

Failure to randomise and/or blind leads to more “positive” results

Blind/not blind	odds ratio	3.4 (95% CI 1.7-6.9)
Random/not random	odds ratio	3.2 (95% CI 1.3-7.7)
Blind Random/ not blind random	odds ratio	5.2 (95% CI 2.0-13.5)

290 animal studies scored for blinding, randomisation and positive/negative outcome, as defined by authors

Bebarta et al 2003 Acad. emerg. med. 10:684-687

Classification variables

Some variables such as gender and genotype are “classifications” instead of being “treatments”.

Animals to be compared should be the same age and from the same environment and should be housed and measured in random order.

Inbred strains or outbred stocks

Isogenic strains (inbred, F1)

- Isogenic (animals identical)
- Homozygous, breed true (not F1)
- Phenotypically uniform
- Defined (quality control)
- Genetically stable
- Extensive background data with genetic profile
- Internationally distributed

Like immortal clones of genetically identical individuals. Several hundred strains available.

Outbred stocks

- Each individual different
- Do not breed true
- Phenotypically variable
- Not defined (no QC)
- Genetic drift can be rapid
- Validity of background data questionable. No genetic profile
- Not internationally distributed

Cheap and widely used, but the cost of the animals is a small proportion of total costs

22 Nobel prizes since 1960 where use of inbred strains was essential

Immunology

Medawar and Burnet- Immunological tolerance (1960)
Doherty and Zinkanage-MHC restriction (1996)
Beutler and Steinman-innate immunity (2011)
Tonegawa-antibody diversity (1987)
Jerne -T-cell receptor (1984)
Snell-Transplantation loci (1980)
Kohler and Milstein-monoclonal antibodies (1984)

Genetic modification

Evans-embryonic stem cells (2007)
Capecchi-homologous recombination (2007)
Smithies-genetic modification (2007)

Genetics

Axel and Buck-genes for olfaction (2004)

Transmissable encephalopathies

Pruisiner (1997)

Growth factors

Cohen, Levi-Montalcini (1986)

Cancer

Varmus (1989), Bishop (1989), Baltimore (1975), Temin (1975)

Why do scientists continue to use outbred stocks when inbred strains are available?

Humans are outbred
We wish to model humans

Therefore we should use outbred animals

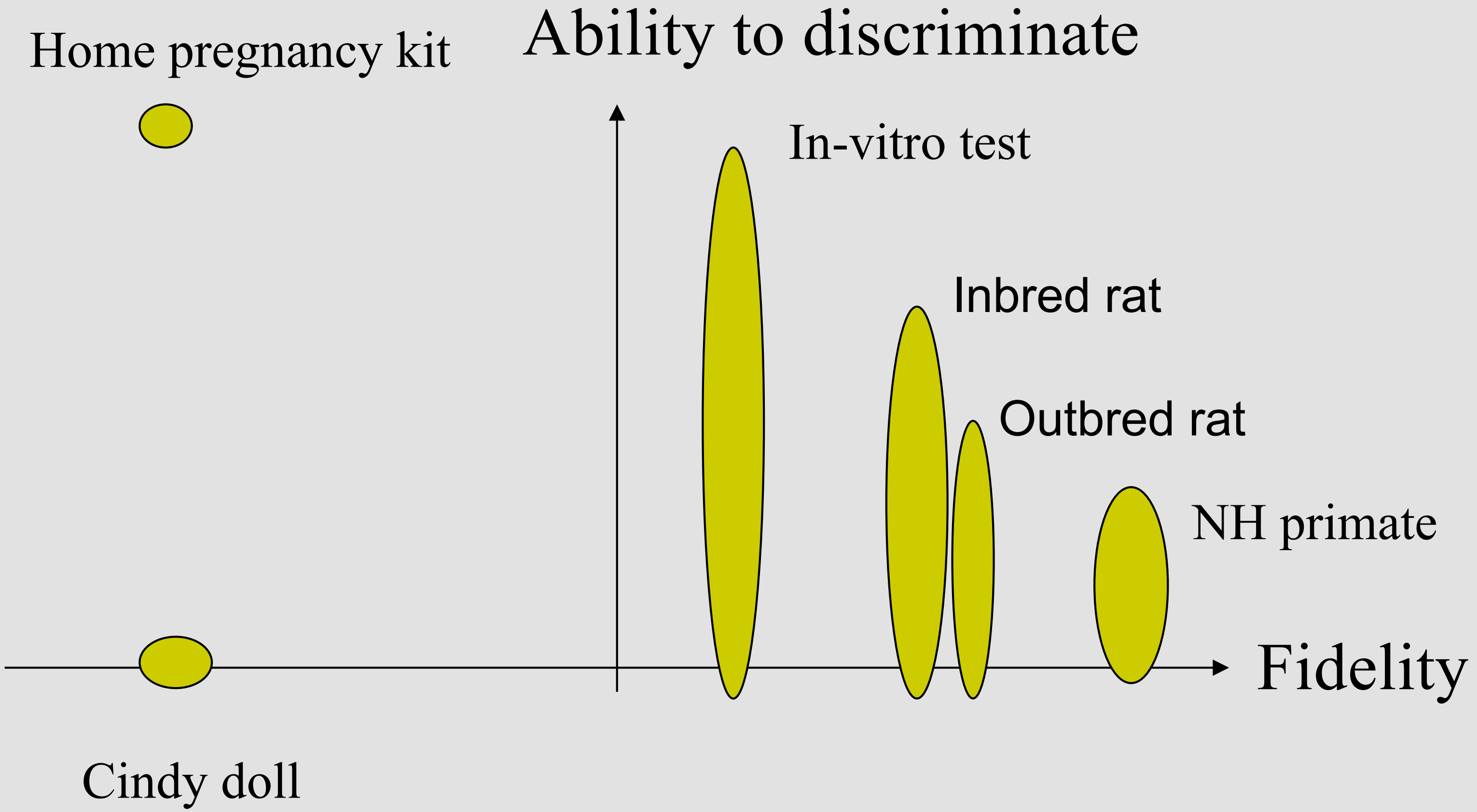
Why do scientists continue to use outbred stocks when inbred strains are available?

Humans weigh 70 kg
We wish to model humans

Therefore we should use 70 kg animals

What do we mean by “model” ?

Models and the high fidelity fallacy (after Russell and Burch)



EU 10.1 (Describe the concepts of fidelity and discrimination (e.g. as discussed by Russell and Burch and others).

The determination of sample size

Three methods of determining sample size

"Tradition"

Copy other investigators in the same discipline. Some merit, but ERCs & funders often want a power analysis.

Resource equation

Based on practical experience. Experiments should have between about 10 and 20 degrees of freedom in the analysis of variance of the results. But ERCs & funders often want a power analysis..

Power analysis

Makes use of the mathematical relationship between the six variables that can determine sample size when there are two treatments. Complex and widely misunderstood. It is not an objective method of determining sample size because it requires a subjective estimate of the minimum effect size likely to be of scientific interest. It also has "spurious precision"

Tradition

“Except in rare instances…., a decision on the size of the experiment is bound to be largely a matter of judgement and some of the more formal approaches to determining the size of the experiment have spurious precision”.

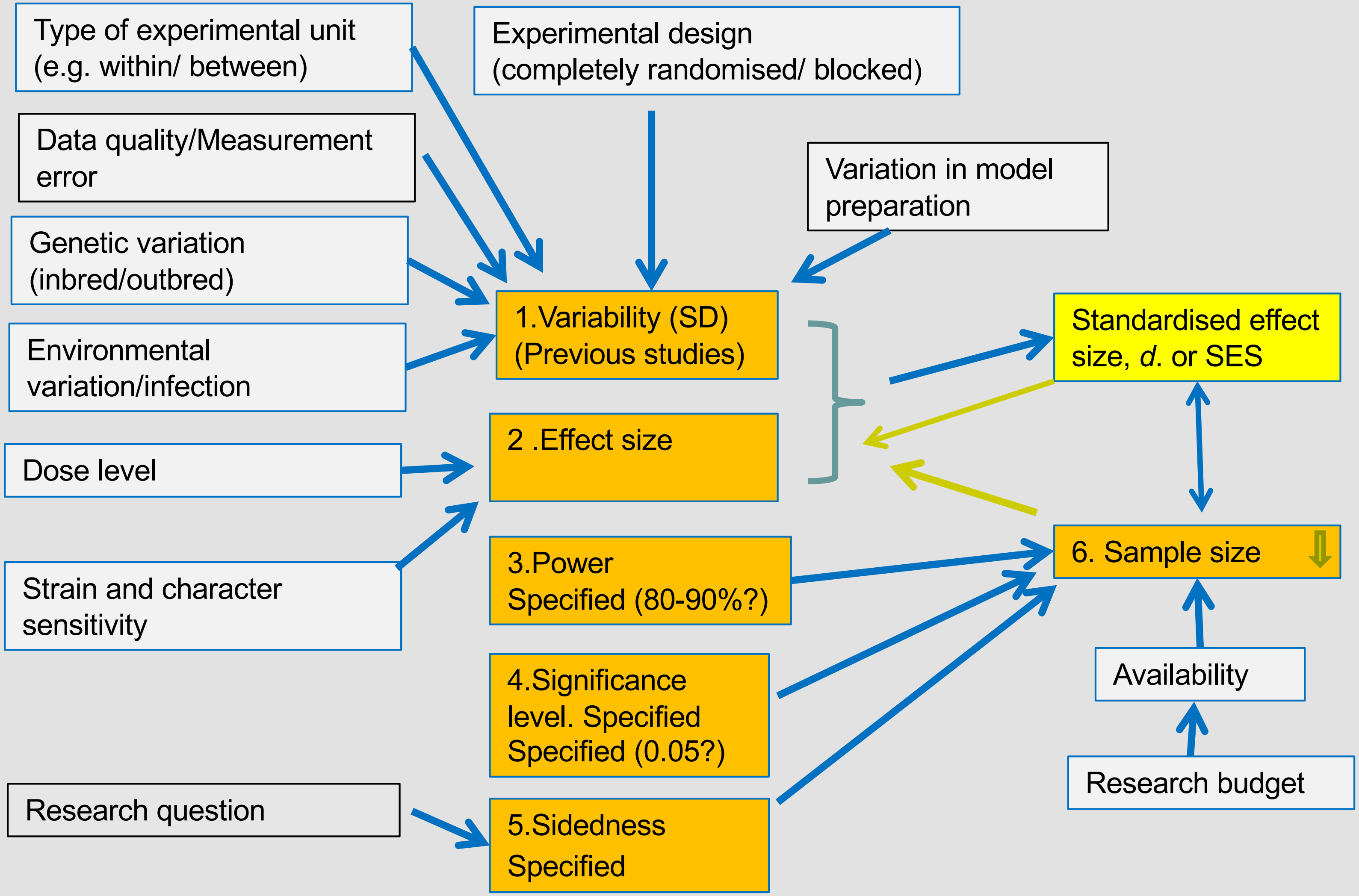
Cox DR, Reid N. The theory of the design of experiments. Boca Raton, Florida: Chapman and Hall/CRC Press; 2000.

Sir David Cox has written two books on experimental design and is the first winner of the “International Statistics prize”. There are few other statisticians in the world who are as highly respected. He and Dr Reid are clearly referring to the power analysis when they mention “spurious precision”

Power Analysis for sample size and effects of variation

- A mathematical relationship between six variables. Fix five of these to determine the 6th one.
- Needs subjective estimate of effect size to be detected (signal)
- Has to be done separately for each character
- Not easy to apply to complex designs
- Essential for expensive, simple, large experiments (clinical trials)
- Useful for exploring effect of variability
- Not objective. It requires an estimate of size of treatment effect that the investigator wants to be able to detect

Factors affecting power and sample size



Type of experimental unit (e.g. within/ between)

Data quality/Measurement error

Genetic variation (inbred/outbred)

Environmental variation/infection

Dose level

Strain and character sensitivity

Research question

Experimental design (completely randomised/ blocked)

Variation in model preparation

1.Variability (SD) (Previous studies)

2 .Effect size

3.Power Specified (80-90%?)

4.Significance level. Specified Specified (0.05?)

5.Sidedness Specified

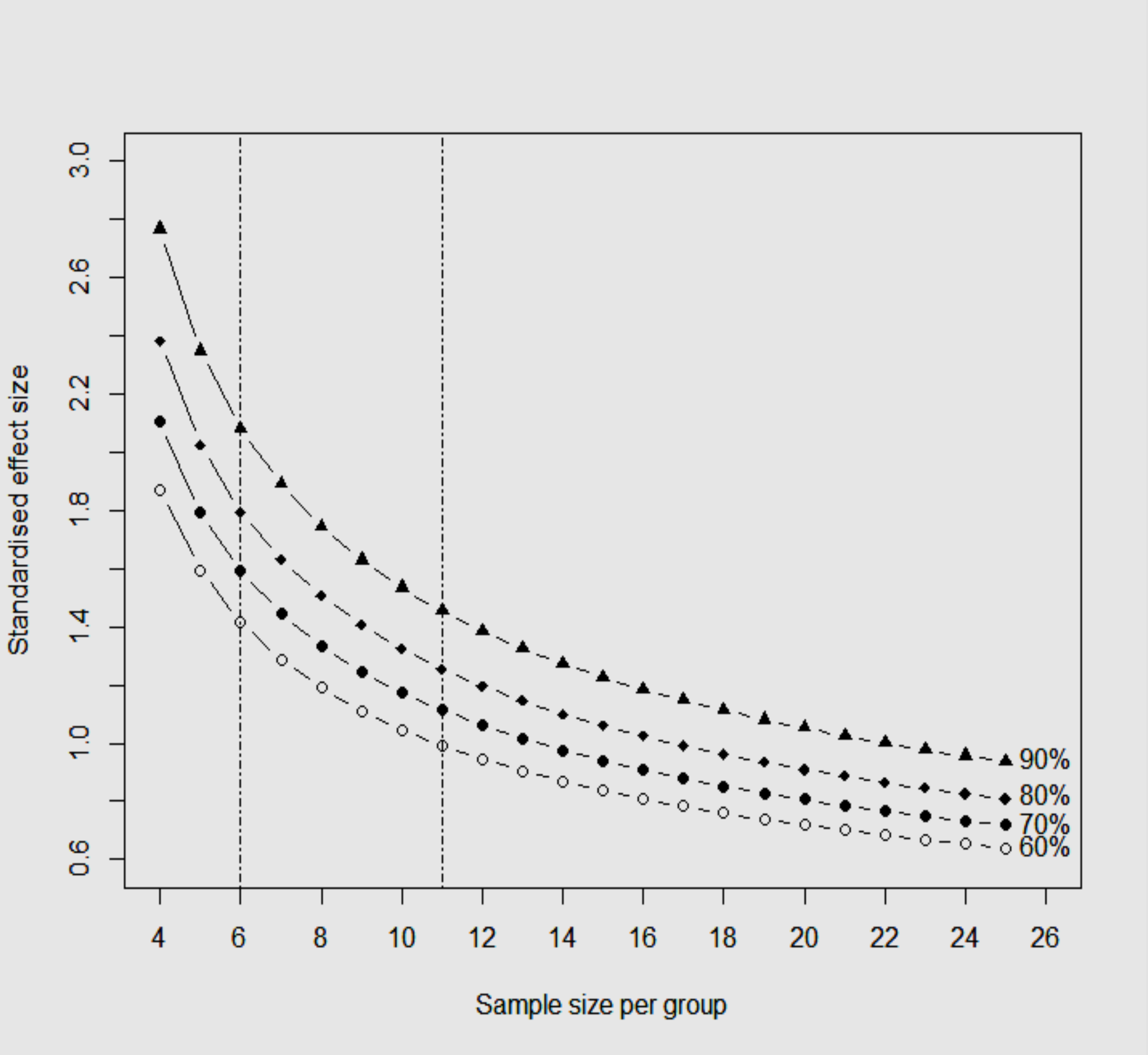
Standardised effect size, *d*. or SES

6. Sample size ⇩

Availability

Research budget

Standardised effect size (d) as a function of sample size for four levels of power



Assuming a 2-sided test.

Vertical lines correspond to sample sizes for the Resource Equation method.

A simplified way of determining sample size using a power analysis.

SES (Cohen's *d*) for 80% & 90% power one or two sided assuming a 5% significance level

Sample size	80% one sided	90% one sided	80% Two-sided	90% Two-sided
4	2.00	2.35	2.38	2.77
5	1.72	2.03	2.02	2.35
6	1.54	1.82	1.80	2.08
7	1.41	1.66	1.63	1.89
8	1.31	1.54	1.51	1.74
9	1.23	1.44	1.41	1.63
10	1.16	1.36	1.32	1.53
11	1.10	1.29	1.26	1.45
12	1.05	1.23	1.20	1.39
13	1.00	1.18	1.15	1.33
14	0.97	1.14	1.10	1.27
15	0.93	1.10	1.06	1.23
16	0.90	1.06	1.02	1.18
17	0.87	1.03	0.99	1.15
18	0.85	1.00	0.96	1.11
19	0.82	0.97	0.93	1.08
20	0.80	0.94	0.91	1.05
21	0.78	0.92	0.89	1.03
22	0.76	0.90	0.86	1.00
24	0.73	0.86	0.83	0.96
26	0.70	0.82	0.79	0.92
28	0.67	0.79	0.76	0.88
30	0.65	0.76	0.74	0.85
32	0.63	0.74	0.71	0.82
34	0.61	0.72	0.69	0.80

Suggested procedure

1. Find an SD for character of interest
2. Choose a sample size based on previous experience/published work, available resources
3. Look in table (left) to find Cohen's *d* for chosen power and sidedness
4. Multiply *d* by the SD to get effect size (ES: difference between means) in original units
5. Decide whether this ES is sufficient. e.g.. would it be better to be able to find a smaller ES? If so, choose a larger sample size and repeat.
6. Explain any calculations and assumptions in manuscript

Estimating sample/effect size for an experiment

Sample Size	d or SES 90% 2 sided
4	2.77
5	2.35
6	2.08
7	1.89
8	1.74
9	1.63
10	1.53
11	1.45
12	**1.39**
13	1.33
14	1.27
15	1.23
16	1.18
17	1.15
18	1.11
19	1.08
20	1.05
21	1.03
22	1.00
24	0.96
26	0.92
28	0.88
30	0.85
32	0.82
34	0.80

Question: **Does your new drug alter RBC count in mice?**

1. From literature C57BL /6 mice have a mean Red Blood Cell count of 9.19, SD=0.40 (n/μL).

2. Say your preliminary choice is a sample size of n=12 mice/group

3. From table, left, for 90% power, two sided d=1.39

4. Detectable effect size=d*SD, = 1.39*0.40=0.56 n/μL

5. This is a (0.56/9.19)*100= 6% change.

6. Is this OK? If not, change sample size.
If you used 24 mice/group the predicted ES would be 0.96*0.40=0.38 n/μL., a 4% change

7. You state: “From published work the mean and standard deviation of RBC in C57BL/6 mice is about 9.19±0.40. Using a power analysis I estimated that a sample size of n=12 will provide a 90% chance of detecting a change in RBC count of 0.56 n/μL or 6%.”

This depends on getting an SD of 0.40 or less

Cohen’s *d* from previous examples

Example 1. Effect of MCA and urethane on micronuclei.
MCA, *d*=0.22 (ns)
Urethane *d*=1.96

Example 2. Apoptosis in rat thymocytes
CPG *d*=1.85(ns)
STAU *d*=3.91

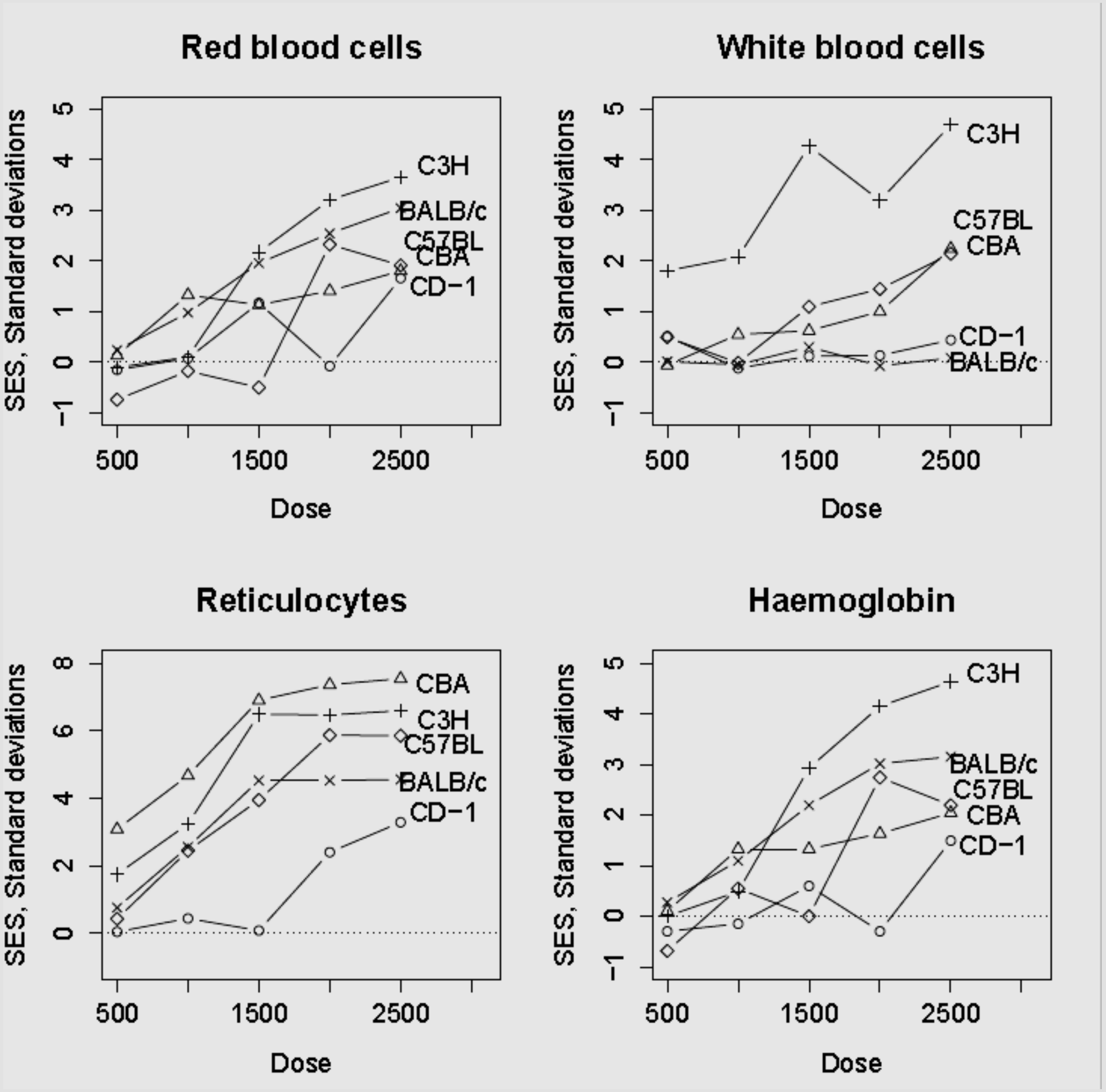
Example 3. Explaining factorial designs (see next slide)

Example 4. Randomised block factorial design, effect of diallyl sulphide
d=4.4

Other studies: Cohen’s d is often well above 2.0 SDs in laboratory animal experiments.

So sample sizes can be small if variation controlled.

Examples of Cohen’s *d* (SES) in chloramphenicol experiment.



d=1 “extra large”,
d=2 “gigantic”

Note differences due to

1. Strain
2. Dose
3. Character

Data from:

Festing MFW, Diamanti P, Turton JA. Strain differences in haematological response to chloramphenicol succinate in mice: implications for toxicological research. Food and Chemical Toxicology 2001;39:375-83.

The ARRIVE Guidelines. Main headings

1. TITLE.

2. ABSTRACT

INTRODUCTION

3. Background.

⟹ 4. Objectives.

METHODS

5. Ethical statement

⟹ 6. Study design

7. Experimental procedures.

8. Experimental animals

9. Housing and husbandry

⟹ 10. Sample size

⟹ 11. Allocating animals to experimental groups

12. Experimental outcomes

⟹ 13. Statistical methods

RESULTS

14. Baseline data

15. Numbers analysed

16. Outcomes and estimation

17. Adverse events

DISCUSSION

18. Interpretation/scientific implications

19. Generalisability/translation.

20 Funding

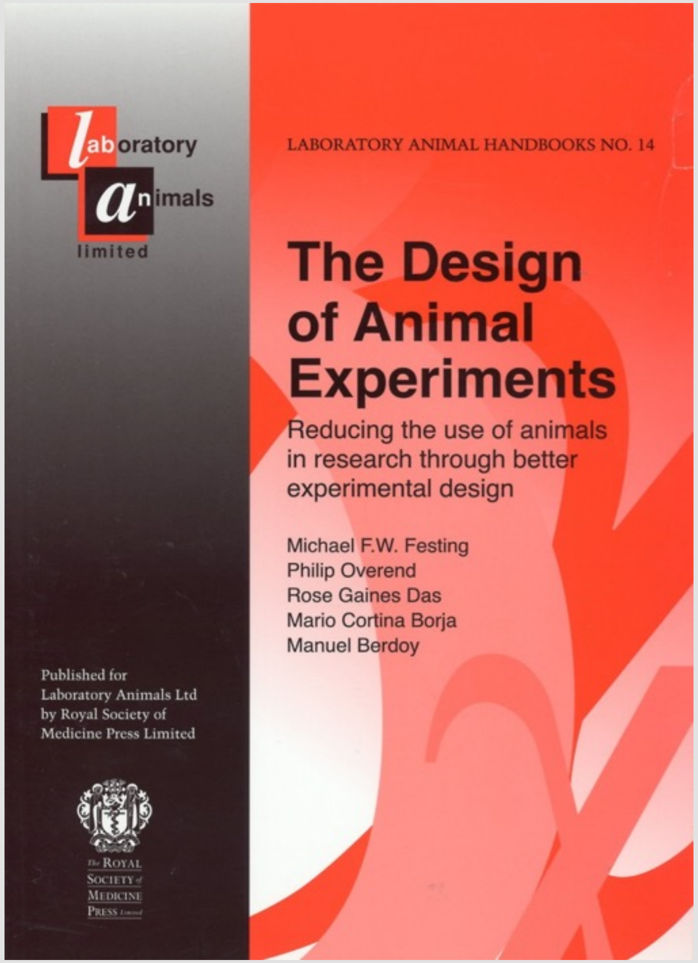
Kilkenny,C., W.J.Browne, I.C.Cuthill, M.Emerson, and D.G.Altman. 2010. "Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research." *PLoS.Biol.* 8:e1000412.

Design of procedures and projects (level 1) – EU Modules 10 and 11

1. Describe the concepts of fidelity and discrimination (e.g. as discussed by Russell and Burch and others).
2. Explain the concept of variability, its causes and methods of reducing it (uses and limitations of isogenic strains, outbred stocks, genetically modified strains, sourcing, stress and the value of habituation, clinical or sub-clinical infections, and basic biology).
3. Describe possible causes of bias and ways of alleviating it (e.g. formal randomisation, blind trials and possible actions when randomisation and blinding are not possible).
4. Identify the experimental unit and recognise issues of non-independence (pseudo- replication).
5. Describe the variables affecting significance, including the meaning of statistical power and “p-values”.
6. Identify formal ways of determining of sample size (power analysis or the resource equation method).
7. List the different types of formal experimental designs (e.g. completely randomised, randomised block, repeated measures [within subject], Latin square and factorial experimental designs).
8. Explain how to access expert help in the design of an experiment and the interpretation of experimental results

Design of procedures and projects (level 2) – EU Modules 9, 10, 11Good scientific practice

1. Describe the principles of a good scientific strategy that are necessary to achieve robust results, including the need for definition of clear and unambiguous hypotheses, good experimental design, experimental measures and analysis of results. Provide examples of the consequences of failing to implement sound scientific strategy
2. Demonstrate an understanding of the need to take expert advice and use appropriate statistical methods, recognise causes of biological variability, and ensure consistency between experiments.
3. Discuss the importance of being able to justify on both scientific and ethical grounds, the decision to use living animals, including the choice of models, their origins, estimated numbers and life stages. Describe the scientific, ethical and welfare factors influencing the choice of an appropriate animal or non-animal model.
4. Describe situations when pilot experiments may be necessary.
5. Explain the need to be up to date with developments in laboratory animal science and technology so as to ensure good science and animal welfare
6. Explain the importance of rigorous scientific technique and the requirements of assured quality standards such as GLP.
7. Explain the importance of dissemination of the study results irrespective of the outcome and describe the key issues to be reported when using live animals in research e.g. ARRIVE guidelines.



laboratory animals limited

LABORATORY ANIMAL HANDBOOKS NO. 14

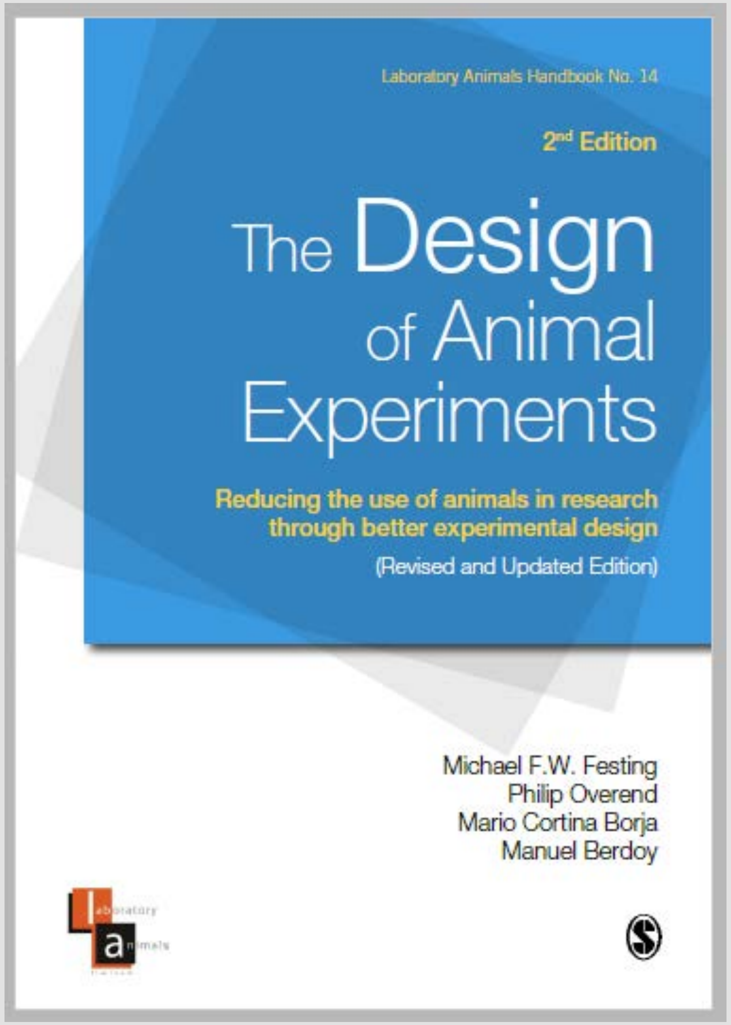
The Design of Animal Experiments

Reducing the use of animals in research through better experimental design

Michael F.W. Festing
Philip Overend
Rose Gaines Das
Mario Cortina Borja
Manuel Berdoy

Published for
Laboratory Animals Ltd
by Royal Society of
Medicine Press Limited

The Royal Society of Medicine Press Limited



Laboratory Animals Handbook No. 14

2nd Edition

The Design of Animal Experiments

Reducing the use of animals in research through better experimental design

(Revised and Updated Edition)

Michael F.W. Festing
Philip Overend
Mario Cortina Borja
Manuel Berdoy

laboratory animals

2002

2016

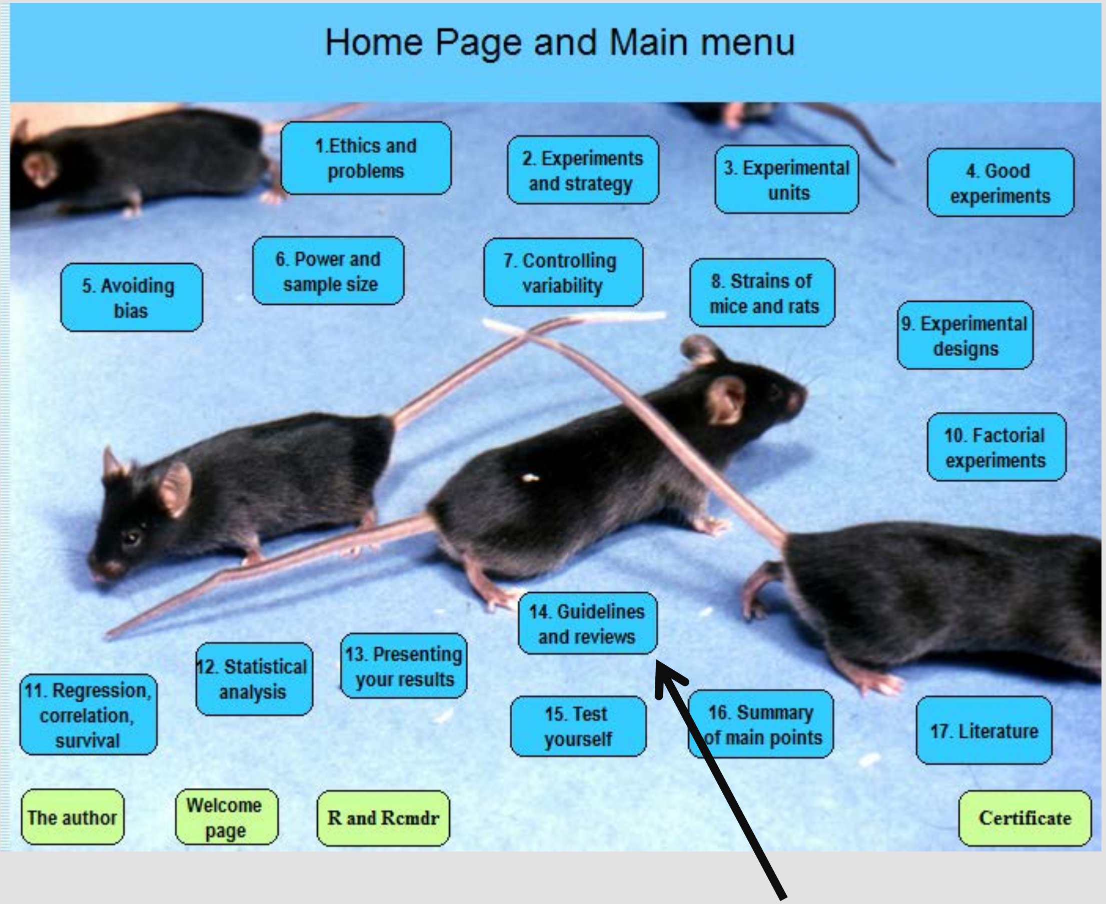
https://uk.sagepub.com/en-gb/eur/the-design-of-animal-experiments/book252408

ISBN: 9781473974630 £15.99

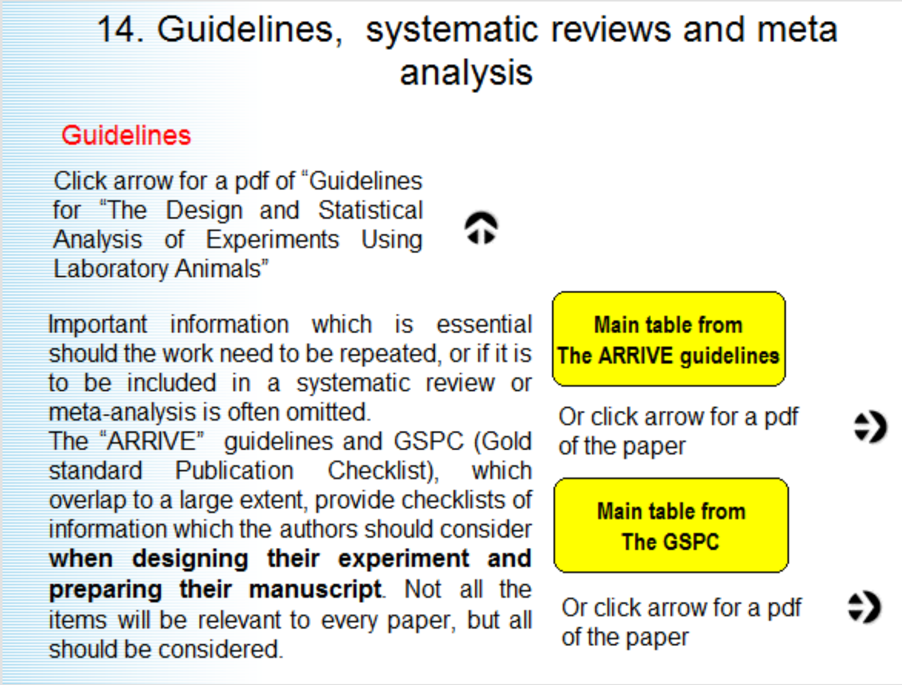
www. 3Rs- Reduction.co.uk

This site provides an interactive short course on experimental design for research scientists working with laboratory animals. The aim is to reduce the number of animals which are used, improve the quality of the science and save time, money and other scientific resources. Ethical review committees, IACUCs and Ph.D. supervisors might like to ask scientists starting work with animals to visit the site, work through it sequentially, and certify (using the form provided) that they have done so before starting their experiments.

www.3Rs-reduction.co.uk



Home Page and Main menu



14. Guidelines, systematic reviews and meta analysis

Guidelines

Click arrow for a pdf of "Guidelines for "The Design and Statistical Analysis of Experiments Using Laboratory Animals"

Important information which is essential should the work need to be repeated, or if it is to be included in a systematic review or meta-analysis is often omitted.
The "ARRIVE" guidelines and GSPC (Gold standard Publication Checklist), which overlap to a large extent, provide checklists of information which the authors should consider **when designing their experiment and preparing their manuscript**. Not all the items will be relevant to every paper, but all should be considered.

Main table from The ARRIVE guidelines

Or click arrow for a pdf of the paper

Main table from The GSPC

Or click arrow for a pdf of the paper

Conclusions

- We are not born knowing how to design a randomised controlled experiment. We need to be taught how to do so.
- Clearly, animal experiments are not always well designed
- Five requirements for a good design
  - Unbiased (randomisation, blinding, randomized block design)
  - Powerful (control variability, uniform materials, blocking)
  - Wide range of applicability, e.g. using factorial designs
  - Simple
  - Amenable to statistical analysis
- Use a power analysis to estimate effect size for a proposed sample size
- Use a randomized block design where possible
- Better training is needed (how?)
- More consultant bio-statisticians should be provided (free?)
- Funding organisations should take responsibility for the quality of the research that they fund!
- Negative results should be as acceptable as positive ones.