Environ Ecol Stat (2014) 21:239–261
DOI 10.1007/s10651-013-0253-4

Prospective evaluation of designs for analysis of variance without knowledge of effect sizes

C. Patrick Doncaster · Andrew J. H. Davey · Philip M. Dixon

Received: 11 September 2011 / Revised: 28 April 2013 / Published online: 21 May 2013

Abstract Estimation of design power requires knowledge of treatment effect size and error variance, which are often unavailable for ecological studies. In the absence of prior information on these parameters, investigators can compare an alternative to a reference design for the same treatment(s) in terms of its precision at equal sensitivity. This measure of relative performance calculates the fractional error variance allowed of the alternative for it to just match the power of the reference. Although first suggested as a design tool in the 1950s, it has received little analysis and no uptake by environmental scientists or ecologists. We calibrate relative performance against the better known criterion of relative efficiency, in order to reveal its unique advantage in controlling sensitivity when considering the precision of estimates. The two measures differ strongly for designs with low replication. For any given design, relative performance at least doubles with each doubling of effective sample size. We show that relative performance is robustly approximated by the ratio of reference to alternative α quantiles of the F distribution, multiplied by the ratio of alternative to reference effective sample sizes. The proxy is easy to calculate, and consistent with exact measures. Approximate or exact measurement of relative performance serves

Handling Editor: Pierre Dutilleul.

C. P. Doncaster (✉)
Centre for Biological Sciences, University of Southampton, Southampton SO17 1BJ, UK
e-mail: cpd@soton.ac.uk

A. J. H. Davey
WRc plc, Frankland Road, Blagrove, Swindon SN5 8YF, UK
e-mail: andrew.davey@wrcplc.co.uk

P. M. Dixon
Department of Statistics, Iowa State University, Ames, IA 50011-1210, USA
e-mail: pdixon@iastate.edu

a useful purpose in enumerating trade-offs between error variance and error degrees of freedom when considering whether to block random variation or to sample from a more or less restricted domain.

Keywords ANOVA mixed models · Experimental design · Power analysis · Sensitivity analysis · Significance test · Statistical power

1 Introduction

Power analysis is used to find the number of observations or the level of background variation that will allow a reasonable probability of detecting a threshold size of effect (Rasch and Herrendörfer 1986; Kraemer and Thiemann 1987). For example, a forester may wish to know whether a fungicide is cost effective. Randomised trials of the treatment against a control can test whether the difference in yield caused by the fungicide is likely to pay for the cost of its application. A good study will choose a sample size that has acceptable power to detect the threshold effect size of interest, which in this case is a difference in yield that is worth as much as the treatment costs. A significant result then tells the investigator that the fungicide is cost effective, within an accepted threshold probability α of making a Type-I error in rejecting a true null hypothesis (often set at 0.05). Alternatively a non-significant result tells the investigator that the fungicide is not cost effective, within an accepted threshold probability β of making a Type-II error in failing to reject a false null hypothesis (where $\beta = 1 - \text{power}$).

The probability of failing to reject a false null hypothesis declines exponentially as a function of sample size (Verrill and Durst 2005), which can create a sharply defined boundary between unsuccessful and successful experiments. Funding councils and research journals increasingly require power calculations to justify the sample sizes of experimental animals or field plots, or other resources. Power analysis is problematic, however, for exploratory studies that have no context for setting a minimum effect size of interest, because the calculation of power requires the unavailable measure of effect size. Power calculation also requires knowledge of the structure and magnitude of error variation, which depend on the choice of study design. For a laboratory experiment, the experimenter may wish to consider alternative treatment procedures for grouping or blocking nuisance variation due to the apparatus. For a field study, the investigator may need to know how the replication between and within sites influences the balance of sensitivities to treatment effects and to regional generality. The aim of this paper is to provide design tools for comparing alternative error structures, which are applicable particularly to exploratory studies with a focus on detecting presence or absence of treatment effects.

Empirical studies often present alternative options for blocking nuisance variation, in the laboratory at the stage of designing experimental facilities, or in the field when seeking candidate sites. Blocking is intended to reduce the error variance, but it also reduces the error degrees of freedom (d.f.). The first effect increases power (assuming normality); the second generally reduces it (though not always monotonically: Blair et al. 1994). Traditional comparisons between designs focus on the relative efficiency of the mean, which evaluates the reduction in error variance achieved by blocking. It provides the appropriate information to choose a study design and sample size when decisions are based on the precision of the mean. The concept of relative performance broadens the scope of relative efficiency by controlling sensitivity in the comparison of two study designs. It compares alternative designs with different error d.f. by computing the change in error variance required to sustain the power of the test. It addresses the question, "if one design is sufficiently powerful to detect a specified treatment effect, will another have at least as good a chance of doing so?". Cochran and Cox (1957) were the first to propose controlling the confidence interval width or the power. Others have compared efficiency at constant power for particular designs (e.g., Abou-el-Fittouh 1976, 1978; Vonesh 1983; Shieh and Show-Li 2004; Wang and Hering 2005); all have had negligible uptake in the ecological and environmental literature, due partly to the difficulty of exact calculation. Despite a resurgent interest in sensitivity analysis (Bacchetti 2010; Lai and Kelley 2012) and in accuracy of parameter estimation (Maxwell et al. 2008; Webb et al. 2010), studies rarely evaluate alternative options for absorbing or controlling error variation in terms of design sensitivity. To date no general analysis and guidance exists for comparing performance at equal power.

Here we provide the first formal comparison of relative performance to relative efficiency, and we develop an easy-to-use proxy for calculating relative performance. We show that the usual adjustment to relative efficiency to account for differences in d.f. is poor at controlling power for designs with few samples and little within-sample replication. Because relative performance explicitly controls power, it is well suited to comparing design options at the planning stage for a study. We consider alternative designs for analysis of variance on the same treatment or treatment combination against a null hypothesis of zero effect. With the same test hypothesis for both designs, we nominate an acceptable level of power against which to evaluate the performance of one design relative to the other in absorbing or controlling random variation within the study population. The resulting measures of relative performance have an advantage over conventional power analysis in permitting objective comparisons without need of predefined sizes for treatment effect and error variance. The price of accommodating this level of ignorance is that designs can only be compared at matching power and cannot be optimised for power. In this article we evaluate the utility of approximate and exact relative performance for comparisons between alternative study designs for the same treatment and population of interest.

1.1 Motivating example

Consider the hypothesis that elevated atmospheric CO_2 has an interactive effect with soil N on growth of poplar seedlings. Suppose that an experimental test of the $CO_2 \times N$ interaction can be done in controlled environment rooms on individually potted seedlings of similar age that sample a population of known source. The test of treatment interactions presents several design options. One is to use 12 rooms, in a fully randomized (FR) allocation of three rooms to each of the four combinations of elevated or ambient N with elevated or ambient CO_2, and r replicate pots in each room. Using $r = 4$ replicates would give this design a power of 0.86 to detect a treatment interaction that has a unitary standardized effect size ($\theta/\sigma = 1$, as defined in the next section) for analysis of variance with $p = 1$ test d.f. and $q = 8$ error d.f. and threshold Type-I error $\alpha = 0.05$. An alternative option is to use six rooms, in a mixed-model split-plot (MM-SP) allocation of three rooms to each level of CO_2. With each room taking r pots at each of elevated and ambient N, this option also uses $12r$ pots in total, and has $q = 4$ error d.f.

In studies of this sort it is common to have no prior effect size of interest or knowledge of the magnitude of error variance. Comparison between alternative designs is nevertheless informed by calculating the relative sizes of error variances for one design to match the other in its power to detect the treatment interaction of interest. For the question of how best to distribute treatments in the controlled-environment study, we can estimate the performance of the MM-SP design relative to the FR at a reasonable power, say of 0.80. We then find that the MM-SP sustains power only if the blocking by room reduces the error variance to $\sim 69\,\%$ of its FR value. We will show that this approximation obtains from $qF(0.95, 1, 8)/qF(0.95, 1, 4) = 5.32/7.71 = 0.69$ where $qF(1 - \alpha, p, q)$ is the α quantile of the F distribution. Conversely, an MM-SP design with ~ 1.45 times more pots can sustain the same error variance as the FR without loss of power, obtained from $qF(0.95, 1, 4)/qF(0.95, 1, 8) = 1.45$. For a conservative expectation of no difference in error variance between the two options, the investigator can now evaluate relative costs and savings of growing $45\,\%$ more seedlings whilst deploying half as many controlled environment rooms. These various scenarios quantify the trade-offs amongst design options that inform planning decisions. If preliminary data can be collected, they inform design of the pilot study.

Field studies often use random factors expressly to investigate the generality of effects at different spatial and temporal scales. In such cases, the type of design is generally fixed by the treatment(s) and in situ population of interest, and design considerations focus on optimising the amount of replication at relevant scales. For example the CO_2-by-N effect on plant growth can be tested in the field on adult trees using CO_2 ring diffusers in forest plots, in which case the mixed-model split-plot design needs a further split to incorporate a site block. It then calibrates the treatment effects against within-treatment variation between sites, between plots within sites, and between trees within plots. Distributing the diffusers across more replicate sites has the advantage of raising the error d.f., though at the likely cost of also raising the error variance. To quantify this trade-off requires a prospective calculation of relative performance on alternative numbers of replicate sites and plots.

The following sections develop the concept of relative performance, define the method of approximating it from ratios of critical F quantiles, and illustrate applications with worked examples.

2 Methods

2.1 Relative efficiency

Relative efficiency is the relative amount of information about the mean provided by a single observation from each of two designs (Neyman et al. 1935; Fisher 1935). For an observation from a random variable with a normal distribution with variance σ^2, the Fisher information is $1/\sigma^2$, so the relative efficiency is the ratio of the error variances (Cochran and Cox 1957; Steel and Torrie 1960). For example, a blocked design such as a randomized split plot has relative efficiency as an alternative to a fully randomized reference design:

$$\text{RE} = \frac{\sigma^2_{ref}}{\sigma^2_{alt}}. \tag{1}$$

Here and throughout the paper, σ^2 refers to the quantity that is estimated in the study by the error mean square due to treatment replication (e.g., of laboratory rooms, or of field sites in the motivating examples), which includes variance components from any nested factors.

Equation (1) does not involve the d.f. associated with each error variance because relative efficiency is based on the information about the mean contributed by a data point from a normal distribution with specified variance. The usual degree of freedom adjustment considers the information provided by a single observation from a random variable with a t-distribution at q d.f. (Fisher 1960, pp. 242–244). That information is $(q+1)/[(q+3)\sigma^2]$, giving an adjusted relative efficiency of:

$$\text{RE}_{\text{adj}} = \frac{\sigma^2_{ref}}{\sigma^2_{alt}} \times \left(\frac{q_{ref}+3}{q_{ref}+1} \times \frac{q_{alt}+1}{q_{alt}+3}\right). \tag{2}$$

2.2 Power calculation for an F test

Statistical power equals $1-\beta$, where β is the type-II error rate of retaining a false null hypothesis. The power of an F test to detect a true effect of a fixed treatment depends on the error variance σ^2, the effective sample size n, and the variability among treatment population means (e.g., Kirk 1982). Specifically, power increases monotonically with the non-centrality parameter, $\lambda = n \cdot \sum^a (\mu_i - \mu)^2/\sigma^2$, where a is the number of treatments, μ_i is the population treatment mean for treatment i, and μ is the average of the a population treatment means. The effective sample size n equals the product of all of the variables contributing to the total d.f. of the model that do not also contribute to the p of the treatment term. For a given λ, p, q and α, power can be calculated directly from a non-central F-distribution using any of the many available computer programs, web applets or statistical tables.

The treatment effect size θ is the square root of the treatment-only variability per degree of freedom (e.g., Lenth 2006): $\theta = \left(\sum^a (\mu_i - \mu)^2/p\right)^{0.5}$. For the simplest case of a two-level treatment, $\theta = (\mu_1 - \mu_2)/\sqrt{2}$.

2.3 Relative performance

For a given treatment and population of interest, we define the performance of an alternative design relative to a reference design as the ratio of expected error variances, $\sigma^2_{alt}/\sigma^2_{ref}$, for which the two designs have the same power. The expected error variances σ^2_{alt} and σ^2_{ref} for alternative and reference designs respectively are each estimated by the design-specific error mean square in the denominator of the F test statistic for the treatment effect. In the motivating example, the MM-SP design had a performance of ~69 % relative to the FR design with the same total number of pots. The MM-SP is therefore a poor alternative if its blocks are expected to reduce the error variance by less than 31 %, and a good choice if the expectation exceeds 31 %. We will show in the Results that using twice as many pots in the MM-SP design doubles its relative performance to 138 %. This will be a good choice if it is expected to have similar error variance to the FR design, and a cost-effective choice if the cost of twice as many pots is outweighed by the saving in using half as many rooms.

Different designs are comparable in terms of relative performance only when reference and alternative options test the same set of hypotheses. This means that they must (i) apply the same fixed treatment(s), and therefore have the same effect size; (ii) allocate treatment levels to the same scale(s) of sampling unit (i.e., rooms and pots in the example of laboratory options, sites and plots in the example of field options); (iii) measure the response from the same population of interest (i.e., genotype or genotype mix, with seedlings for the laboratory experiment randomly sampled from a definable source, or trees for the field experiment sampled by the random site variable from across a definable region).

Relative performance is cumbersome to calculate because it requires finding the error variances that give reference and alternative design equal power. We therefore present an approximation of relative performance that is easily calculated from standard tables. The approximate relative performance is given by the ratio of critical F quantiles for reference and alternative models, weighted by the ratio of effective sample sizes:

$$\mathrm{RP}_{F\text{-approx}} = \frac{qF\left(1-\alpha,\, p,\, q_{ref}\right)}{qF\left(1-\alpha,\, p,\, q_{alt}\right)} \times \frac{n_{alt}}{n_{ref}}. \tag{3}$$

Here $qF(1-\alpha, p, q)$ is the α quantile of the F distribution with p numerator d.f. and q denominator d.f., readily available from tables or statistical programs. The derivation of the approximation is given in the "Appendix". Its use of the F distribution means that it applies only to analyses with homogeneous variances across samples and a normal distribution of each error term.

We evaluate the quality of this approximation for a range of designs by comparing $\mathrm{RP}_{F\text{-approx}}$ to its exact equivalent given by the non-central F distribution. The applications Piface (Lenth 2006), G*Power 3 (Faul et al. 2007) and R (R Development Core Team 2010) were used to identify design-specific non-centrality parameters, λ, for a range of power values. These yielded the error fractions $\sigma^2_{alt}/\sigma^2_{ref}$ for which two designs have precisely the same power. These calculations of power use a type-I error rate $\alpha = 0.05$. A smaller value such as $\alpha = 0.01$ lowers performances of alternative relative to reference designs. The reduction is uniform across magnitudes of β, however, with the result that the $\mathrm{RP}_{F\text{-approx}}$ is no less effective at lower α. Computer program Performance calculates $\mathrm{RP}_{F\text{-approx}}$ from inputs of n and test and error d.f. It is available from: http://www.personal.soton.ac.uk/cpd/anovas/datasets/Performance.exe.

Comparisons of relative performance will be illustrated with two commonly used alternatives to fully randomized (FR) designs: (i) randomized complete block (CB), and (ii) split-plot (SP), either of which can be fully replicated in a mixed model (MM). We consider balanced designs of one-way treatment structures and of two-way complete factorial treatment structures. Figure 1 illustrates the layout of each of the principal designs, and Table 1 summarizes their models for analysis of variance. Although the model comparisons for two-factor designs focus on treatment interactions, relative performance applies equally to main effects. The method can anticipate model simplification by pooling of error terms, though these post hoc analyses are not considered in our examples.

3 Results

3.1 Comparison of relative efficiency and relative performance

Reference and alternative designs have the same Fisher-adjusted relative efficiency when (from Eq. 2):

$$\frac{\sigma^2_{ref}}{\sigma^2_{alt}} \times \left(\frac{q_{ref}+3}{q_{ref}+1} \times \frac{q_{alt}+1}{q_{alt}+3} \right) = 1. \tag{4}$$

The same two designs will have approximately the same power to detect a specified treatment effect when (from Eq. 3):

$$\frac{\sigma^2_{ref}}{\sigma^2_{alt}} \times \left(\frac{qF\left(1-\alpha, p, q_{ref}\right)}{qF\left(1-\alpha, p, q_{alt}\right)} \times \frac{n_{alt}}{n_{ref}} \right) = 1. \tag{5}$$

The two approaches can be compared in two ways: (i) Compare their adjustment factors, i.e. the quantities within the outer parentheses of Eqs. 4 and 5, for two designs across a range of treatment levels and effective sample sizes; (ii) Use the Fisher degrees-of-freedom adjustment to compute the σ^2_{alt} that gives the two designs the same relative efficiency, find the non-centrality parameter that provides a specified power for the reference design, then use that σ^2_{alt} and non-centrality parameter to find the power of the alternative design. If Fisher-adjusted relative efficiency and relative performance are similar, the two adjustment factors computed in approach (i) will be similar and the power computed in approach (ii) will be similar to the specified power for the reference design.

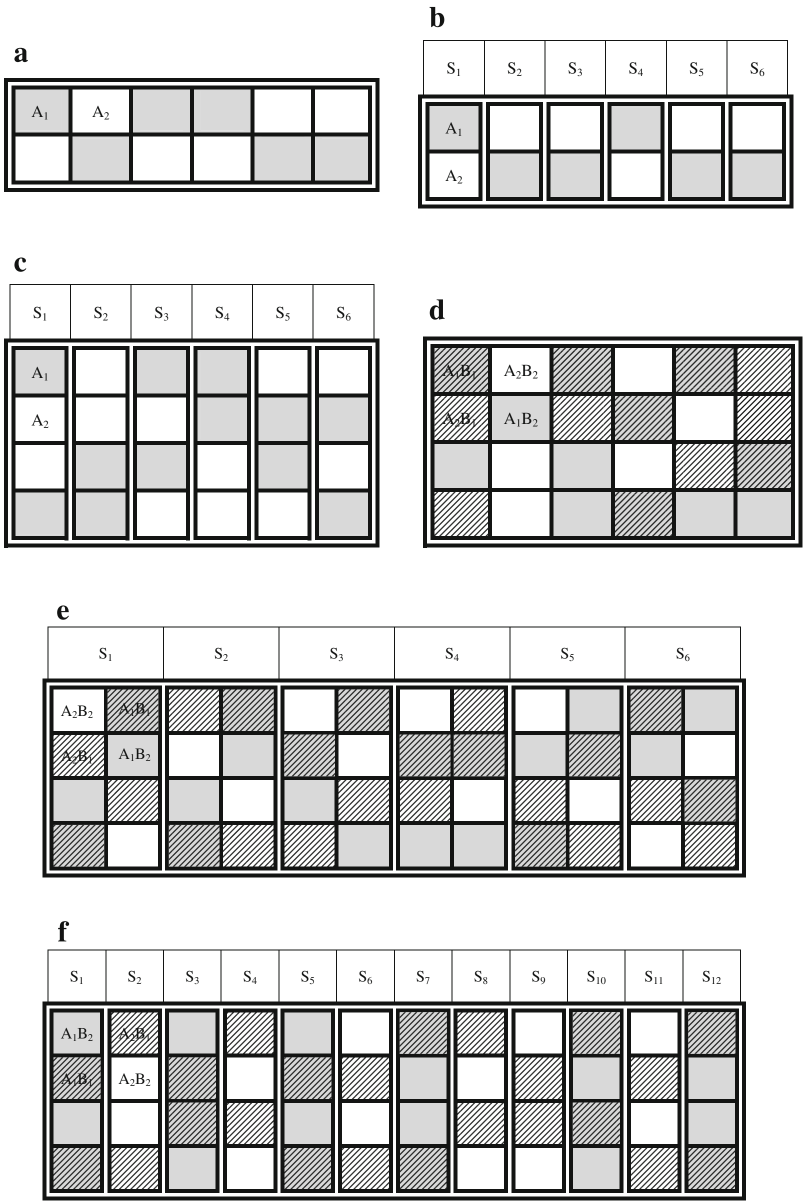

Fig. 1 Example layouts of spatial designs for analysis of variance, with *each cell* representing an observation on a sampling unit. All have $s = 6$ treatment replicates, in (**a**) to (**c**) at each of $a = 2$ levels for testing treatment factor A; in (**d**) to (**f**) at each of $b{\cdot}a = 4$ combinations of levels of treatment factors A and B for testing the B × A interaction. *Double lines* surround a set of sampling units with a randomized allocation of treatments. *Grey*: A_1, *white*: A_2; *hatched lines*: B_1, *no hatching*: B_2. **a** One factor FR design $S'_6(A_2) : Y = A + S'(A)$. **b** One factor CB design $S'_6|A_2 : Y = S' + A + S' \times A$. **c** One factor MM-CB design $R'_2(S'_6|A_2) : Y = S' + A + S' \times A + R'(S' \times A)$. **d** Two factor FR design $S'_6(B_2|A_2) : Y = A + B + B \times A + S(B \times A)$. e Two factor MM-CB design $R'_2(S'_6|B_2|A_2) : Y = A + B + B \times A + S' \times A + S' \times B + S' \times B \times A + R(S' \times B \times A)$. **f** Two factor MM-SP design $R'_2(B_2|S'_6(A_2)) : Y = A + S'(A) + \mathbf{B} + \mathbf{B} \times S'(A) + R'(B \times S'(A))$

Table 1 Five common designs for analysis of variance, showing the model structure, the treatment term and its p d.f., the error term and its q d.f., and the effective sample size n

Design	Structure[a]	Treatment	p	Error	q	n
One treatment factor						
FR: Fully randomized, nested if $r > 1$	$\mathrm{R}'_r(\mathrm{S}'_s(\mathrm{A}_a))$	A	$a - 1$	$\mathrm{S}'(\mathrm{A})$	$(s - 1)a$	$r{\cdot}s$
CB: Randomized complete block, mixed model if $r > 1$	$\mathrm{R}'_r(\mathrm{S}'_s \mid \mathrm{A}_a)$	A	$a - 1$	$\mathrm{S}' \times \mathrm{A}$	$(s - 1)(a - 1)$	$r{\cdot}s$
Two treatment factors						
FR: Fully randomized, nested if $r > 1$	$\mathrm{R}'_r(\mathrm{S}'_s(\mathrm{B}_b \mid \mathrm{A}_a))$	A	$a - 1$	$\mathrm{S}'(\mathrm{B} \times \mathrm{A})$	$(s - 1)b{\cdot}a$	$r{\cdot}s{\cdot}b$
		B	$b - 1$	$\mathrm{S}'(\mathrm{B} \times \mathrm{A})$	$(s - 1)b{\cdot}a$	$r{\cdot}s{\cdot}a$
		B × A	$(b - 1)(a - 1)$	$\mathrm{S}'(\mathrm{B} \times \mathrm{A})$	$(s - 1)b{\cdot}a$	$r{\cdot}s$
CB: Randomized complete block, mixed model if $r > 1$	$\mathrm{R}'_r(\mathrm{S}'_s \mid \mathrm{B}_b \mid \mathrm{A}_a)$	A	$a - 1$	$\mathrm{S}' \times \mathrm{A}$	$(s - 1)(a - 1)$	$r{\cdot}s{\cdot}b$
		B	$b - 1$	$\mathrm{S}' \times \mathrm{B}$	$(s - 1)(b - 1)$	$r{\cdot}s{\cdot}a$
		B × A	$(b - 1)(a - 1)$	$\mathrm{S}' \times \mathrm{B} \times \mathrm{A}$	$(s - 1)(b - 1)(a - 1)$	$r{\cdot}s$
SP: Split plot, mixed model if $r > 1$	$\mathrm{R}'_r(\mathrm{B}_b \mid \mathrm{S}'_s(\mathrm{A}_a))$	A	$a - 1$	$\mathrm{S}'(\mathrm{A})$	$(s - 1)a$	$r{\cdot}s{\cdot}b$
		B	$b - 1$	$\mathrm{B} \times \mathrm{S}'(\mathrm{A})$	$(b - 1)(s - 1)a$	$r{\cdot}s{\cdot}a$
		B × A	$(b - 1)(a - 1)$	$\mathrm{B} \times \mathrm{S}'(\mathrm{A})$	$(b - 1)(s - 1)a$	$r{\cdot}s$

[a] The descriptor of design terms follows the convention of a prime to denote a random factor, open bracket for 'nested in levels of', and vertical separator for 'cross-factored with levels of'

Table 2 Comparison between adjusted relative efficiency (RE_{adj}) and approximate relative performance ($\mathrm{RP}_{F\text{-approx}}$) for a FR reference design $\mathrm{R}'_r(\mathrm{S}'_s(\mathrm{A}_a))$ and a CB alternative design $\mathrm{R}'_r(\mathrm{S}'_s|\mathrm{A}_a)$, both having the same number of treatment levels a and treatment replication s

s	a. Adjustment factor[a]		b. Power of CB at $\mathrm{RE}_{adj} = 1$[b]	
	RE_{adj}	$\mathrm{RP}_{F\text{-approx}}$	$r_{alt}/r_{ref} = 1$	$r_{alt}/r_{ref} = 0.5$
Treatment levels $a = 2$				
3	0.840	$0.416 \times r_{alt}/r_{ref}$	0.5818	0.3697
5	0.873	$0.690 \times r_{alt}/r_{ref}$	0.7275	0.4547
20	0.956	$0.935 \times r_{alt}/r_{ref}$	0.7967	0.5057
100	0.990	$0.988 \times r_{alt}/r_{ref}$	0.7999	0.5086
Treatment levels $a = 3$				
3	0.918	$0.741 \times r_{alt}/r_{ref}$	0.7120	0.4317
5	0.944	$0.871 \times r_{alt}/r_{ref}$	0.7659	0.4626
20	0.984	$0.974 \times r_{alt}/r_{ref}$	0.7955	0.4840
Treatment levels $a = 5$				
3	0.967	$0.906 \times r_{alt}/r_{ref}$	0.7639	0.4467
5	0.980	$0.953 \times r_{alt}/r_{ref}$	0.7829	0.4553
20	0.995	$0.990 \times r_{alt}/r_{ref}$	0.7967	0.4647
Treatment levels $a = 10$				
3	0.991	$0.974 \times r_{alt}/r_{ref}$	0.7857	0.4377
5	0.995	$0.987 \times r_{alt}/r_{ref}$	0.7922	0.4392
20	0.999	$0.997 \times r_{alt}/r_{ref}$	0.7981	0.4428

[a] Adjustment factors are the quantities in parentheses in main-text Eq. (4) for RE_{adj}, and Eq. (5) for $\mathrm{RP}_{F\text{-approx}}$ (at $\alpha = 0.05$ and given $n = r{\cdot}s$)

[b] The power of the alternative design at an adjusted relative efficiency of 1, computed for a reference power of 0.80 at $\alpha = 0.05$

Table 2 gives results for a randomized complete block relative to a fully randomized design with the same 2–10 treatment levels and treatment replication varying from 3 to 100. Table 2(a) shows that the Fisher adjustment factor is always closer to 1.0 than is the relative performance adjustment. The two values are close to each other and close to 1.0 for large numbers of treatment levels and for high treatment replication, and when $r_{alt} = r_{ref}$. Table 2(b) shows that the power of an alternative design with relative efficiency = 1 is less than the reference power of 0.80 when $r_{alt} = r_{ref}$. It is much less when both designs have low treatment replication (e.g., alternative power = 0.58 for 2 treatments and 3 blocks) or whenever $r_{alt} < r_{ref}$. More extreme disparities are seen for $\alpha = 0.01$ (data not shown). These differences reinforce the point that "the ERE [estimated relative efficiency] speaks only to the question of estimation, i.e. precision of estimates, and not to the question of power, i.e. sensitivity of the experiment" (Hinkelmann and Kempthorne 1994, p. 262).

3.2 Relative performance for alternative designs

Here we compare relative performance for situations where the investigator has options amongst two or more different designs to test fixed treatment effects on a population of interest. This is often the case in laboratory manipulations of samples from a known population, where decisions must be made about cost-effective ways to block nuisance variation due to the experimental conditions.

Figure 2 illustrates relative performances for three blocked designs as alternatives to FR references with the same effective sample size, n. The relative performance of the alternative design is calculated at precisely equal power (coloured dots), and approximated from ratios of critical F quantiles with $\mathrm{RP}_{F\text{-approx}}$ (open circles). For both CB and SP alternatives to FR designs, $\mathrm{RP}_{F\text{-approx}}$ closely tracks the precise relative performance at power = 0.8 with low replication, increasing to 0.99 and higher with more replication.

In general, $\mathrm{RP}_{F\text{-approx}}$ slightly overvalues the relative performance of blocks in sustaining high power at large p, and the more so for designs with small n. The 3×3 CB alternative to the FR illustrates this in Fig. 2b, where the middle open circle for $a = b = 3$ (so $p = 4$) and $s = 2$ shows $\mathrm{RP}_{F\text{-approx}}$ corresponding to the error fraction required to match a power of only 0.5 (middle red dot). Nevertheless, its approximation of 57 % for this error fraction is little greater than the 53 % allowed to match a power of 0.8 (middle green dot). For the 5×5 CB and SP designs ($p = 16$), $\mathrm{RP}_{F\text{-approx}}$ slightly overvalues relative performance for any equal powers (except power = α; Fig. 2a–c upper open circles), though negligibly so at high treatment replication. For these designs with high test d.f., particularly when combined with low n, a more conservative proxy is given by the ratio $[qF(1-\alpha, p, q_{ref}) - 1]/[qF(1-\alpha, p, q_{alt}) - 1]$. The "Appendix" gives the rationale for using this adjusted $\mathrm{RP}_{F\text{-approx}}$ for $p > 10$.

To illustrate the application of relative performance, consider the motivating example of a two factor experiment to be carried out in controlled environment rooms. The reference is a nested-FR design: $\mathrm{R}'_r(\mathrm{S}'_s(\mathrm{B}_b|\mathrm{A}_a))$ using the nomenclature of Table 1 and Fig. 1. The alternative is a MM-SP design: $\mathrm{R}'_r(\mathrm{B}_b|\mathrm{S}'_s(\mathrm{A}_a))$. With treatments A and B taking two levels, so $a = b = 2$, and treatment replication of $s = 3$ rooms, the two designs have 8 and 4 error d.f. respectively for testing the interaction. If both have the same r replicate pots, and therefore the same effective sample size $n = r{\cdot}s$ the $\mathrm{RP}_{F\text{-approx}}$ is $qF(0.95, 1, 8)/qF(0.95, 1, 4) = 5.32/7.71 = 0.69$, shown in Fig. 2c by the lower open circle at {3, 0.69}. Accordingly, exact power analyses stipulate that the SP design must have an error variance 69–72 % that of the FR to match reference powers of 0.99–0.80, shown in Fig. 2c by the lower blue dot at {3, 0.69} and lower green dot at {3, 0.72}. If the designs have different r, these relative performances change by r_{alt}/r_{ref}. Though not graphed, the $\mathrm{RP}_{F\text{-approx}}$ of Eq. 3 applies also to comparisons of different s. For example, a MM-SP using $s = 4$ rooms as an alternative to a 3-room FR design (with the same r) has $\mathrm{RP}_{F\text{-approx}} = qF(0.95, 1, 8)/qF(0.95, 1, 6){\times}4/3 = 1.18$.

If we set a threshold of equal error variances between the two designs, then relative performance takes a value of 1 and we can explore options for increasing the effective sample size to sustain 80 % power. This is done by rearranging Eq. (3):

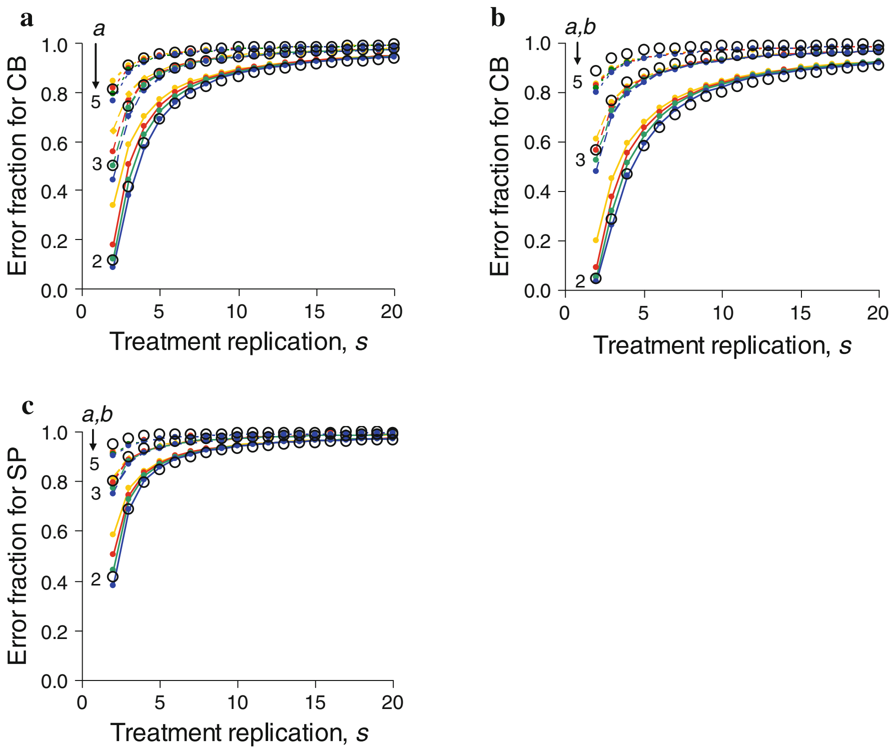

Fig. 2 Relative performance at $\alpha = 0.05$, measured as the error fraction $\sigma^2_{alt}/\sigma^2_{ref}$ required of an alternative design to equal the power $(1-\beta)$ of a fully randomized reference design to detect the same treatment with the same effective samples size $n = r{\cdot}s$. The three groups of lines per graph show results for A in (**a**), and B × A in (**b**–**c**) with 2, 3 and 5 treatment levels; within each group, lower to upper line shows power = 0.99 (*blue*), 0.8 (*green*), 0.5 (*red*), 0.2 (*gold*). *Open circles* are the approximate relative performance $\mathrm{RP}_{F\text{-approx}} = qF(1-\alpha, p, q_{ref})/qF(1-\alpha, p, q_{alt})$. For designs with different r, relative performances must be multiplied by r_{alt}/r_{ref}. For example, the performance of $\mathrm{R}'_3(\mathrm{S}'_5|\mathrm{A}_2)$ relative to $\mathrm{S}'_5(\mathrm{A}_2)$ at power = 0.8 is three times the exact value of 0.72 showing in (**a**), and the corresponding $\mathrm{RP}_{F\text{-approx}}$ is three times the value showing of 0.69. **a** CB design $\mathrm{R}'_r(\mathrm{S}'_s|\mathrm{A}_a)$ versus FR design $\mathrm{R}'_r(\mathrm{S}'_s(\mathrm{A}_a))$. **b** CB design $\mathrm{R}'_r(\mathrm{S}'_s|\mathrm{B}_b|\mathrm{A}_a)$ versus FR design $\mathrm{R}'_r(\mathrm{S}'_s(\mathrm{B}_b|\mathrm{A}_a))$. **c** SP design $\mathrm{R}'_r(\mathrm{B}_b|\mathrm{S}'_s(\mathrm{A}_a))$ versus FR design $\mathrm{R}'_r(\mathrm{S}'_s(\mathrm{B}_b|\mathrm{A}_a))$

$$\frac{n_{alt}}{n_{ref}} = \frac{qF\,(1-\alpha,\, p,\, q_{alt})}{qF\left(1-\alpha,\, p,\, q_{ref}\right)}, \quad \text{at } \mathrm{RP}_{F\text{-approx}} = 1. \tag{6}$$

In the motivating example, each design has its own $n = r{\cdot}s$, and both use $s = 3$ rooms. With $n_{alt}/n_{ref} = qF(0.95, 1, 4)/qF(0.95, 1, 8) = 1.45$, the MM-SP design sustains power by having 1.45 times more pots.

Figure 2 permits generic comparisons of alternative blocking designs. The difference between blocking by CB and by SP diminishes with more replicates, as all approach 100 % performance. A high enough replication will therefore yield power advantages to any kind of blocking even when blocks are likely to absorb little background variation. This result is consistent with simulations designed to explore the value of blocking (Legendre et al. 2004). The relative performance of the SP (or

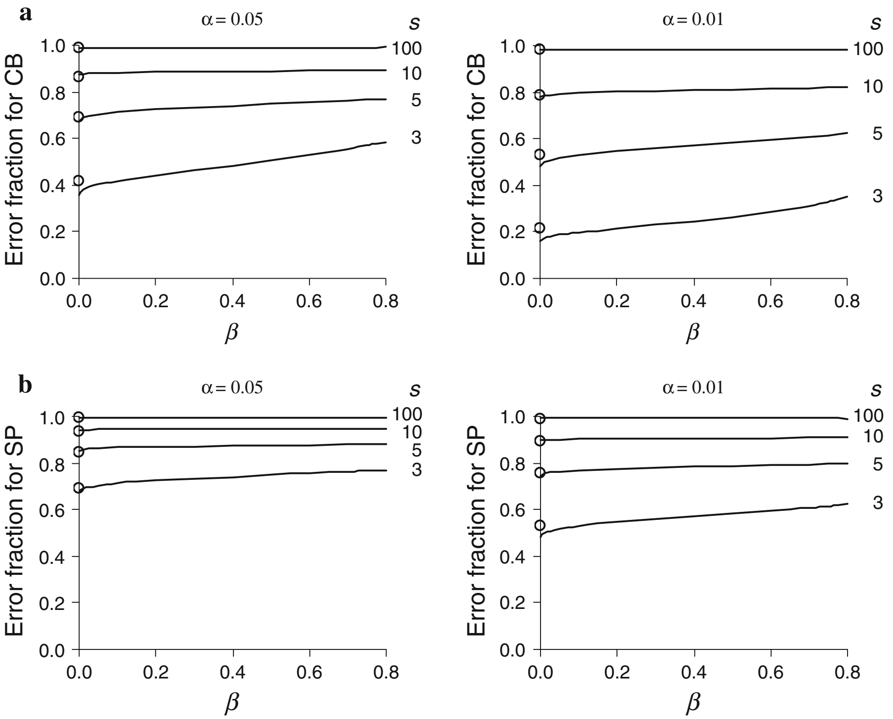

Fig. 3 Error fractions required of CB and SP designs to equal the β (i.e., 1−power) of a fully randomized reference with the same $n = r \cdot s$, for (**a**) A and (**b**) B × A. For each of the four values of s, *open circles* on the vertical axes show $\mathrm{RP}_{F\text{-approx}} = qF(1-\alpha, p, q_{ref})/qF(1-\alpha, p, q_{alt})$. **a** CB design $\mathrm{R}'_r(\mathrm{S}'_s|\mathrm{A}_2)$ versus FR design $\mathrm{R}'_r(\mathrm{S}'_s(\mathrm{A}_2))$. **b** SP design $\mathrm{R}'_r(\mathrm{B}_2|S'_s(\mathrm{A}_2))$ versus FR design $\mathrm{R}'_r(\mathrm{S}'_s(\mathrm{B}_2|\mathrm{A}_2))$

MM-SP) design nevertheless always exceeds the relative performance of the CB (or MM-CB) design for detection of treatment interactions when using unpooled error terms in the analysis of the block design. The intuitive explanation of this difference is that the SP design provides the same or more error degrees of freedom to test an effect than does the CB design with unpooled error terms (Table 1). Although all designs perform equally well given sufficient treatment replication, the CB design for a 2×2 treatment interaction approaches perfect relative performance more slowly than other designs. Even at $s = 20$ blocks, its blocking factor must absorb 8 % of error variance to achieve the power of an SP blocking factor that absorbs only 3 % (Fig. 2b compared to c). When error terms are pooled in the analysis of blocked designs, the relative performance of the CB design is always better than that of the SP design.

The tight banding of coloured lines in Fig. 2 indicates a flat response of performance to power. Figure 3 illustrates for two designs how the error fraction changes little across levels of power, and $\mathrm{RP}_{F\text{-approx}}$ closely matches it at high power. We have further analysed a wide range of other designs at $\alpha = 0.05$ and 0.01, including main effects and interactions in Latin squares, split-split-plot O'(B|R'(S'|A)), and split-split-split-plot C|O'(B|R'(S'|A)) designs with and without pooling of error terms. All

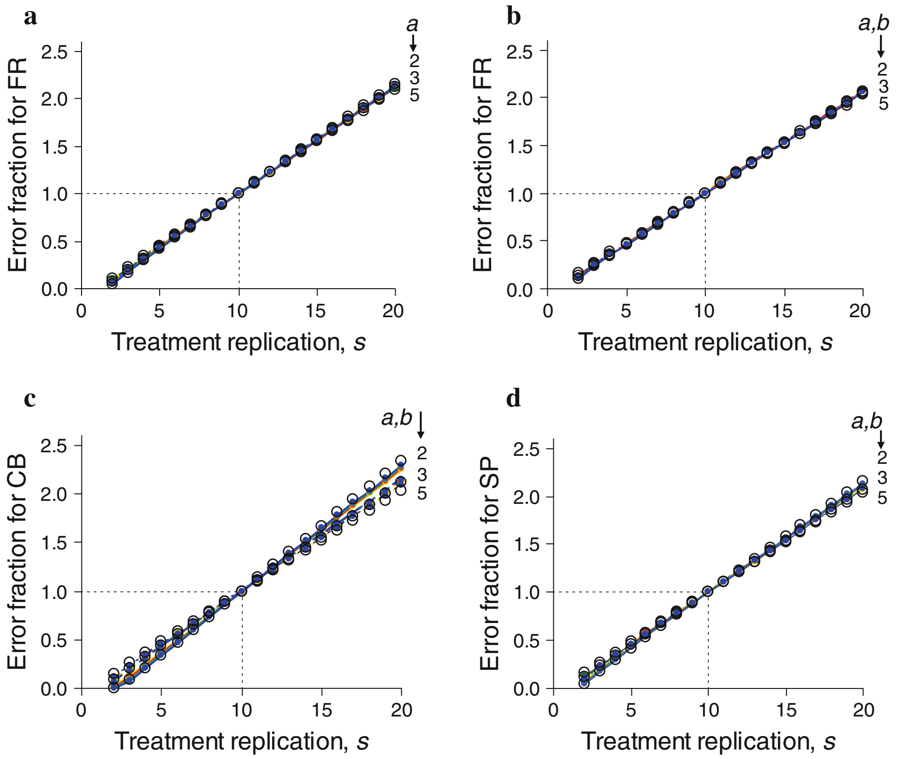

Fig. 4 Relative performance at $\alpha = 0.05$ of a design with alternative treatment replication s to a reference $s = 10$ for the same design. Colour coding and factor levels as for Fig. 2, with lower β overlying higher β. Error fractions assume equal r; otherwise multiply responses by r_{alt}/r_{ref}. *Open circles* are the $\text{RP}_{F\text{-approx}} = qF(1-\alpha, p, q_{ref})/qF(1-\alpha, p, q_{alt}) \times n_{alt}/n_{ref}$, here with $n_{alt}/n_{ref} = s/10$. **a** FR design $\text{S}'_s(\text{A}_a)$. **b** FR design $\text{S}'_s(\text{B}_b|\text{A}_a)$. **c** CB design $\text{R}'_r(\text{S}'_s|\text{B}_b|\text{A}_a)$. **d** SP design $\text{R}'_r(\text{B}_b|\text{S}'_s(\text{A}_a))$

have features entirely consistent with those in Figs. 2 and 3: little variation in relative performance across β for a given ratio of n, and the alternative design achieving the power of a high-powered reference with an error fraction that is closely matched by the proxy, particularly at low p or high s.

3.3 Optimising the trade-off between the number of replicates and their homogeneity

Here we consider experimental or mensurative tests that involve investigating the generality of fixed treatment effects across random spatial or temporal variation, which generally means that the design is set by the test question. For a chosen design, the investigator wishes to compare relative performance with different amounts of replication at relevant scales. This is often the case in field studies where replicates take up space and therefore tend to be less homogeneous when there are more of them.

For treatment replication above or below a reference $s = 10$, Fig. 4 shows that the criterion for equalling power is met with an error fraction that is almost identical across powers. It generally increases in linear proportion to the sample size. Thus a FR design with sample size $s = 20$ instead of 10 sampling units accommodates a little over twice the error variance without loss of power, regardless of the power of the less-replicated reference; conversely, a design with $s = 5$ equals the power achieved by the reference design of $s = 10$ with just less than half the error variance (Fig. 4a, b). All have the threshold error fraction accurately predicted by $\text{RP}_{F\text{-approx}}$ (Fig. 4 open circles). More generally for any design with s-dependent q, an alternative s twice the size of any reference s gives (from Eq. 3): $\text{RP}_{F\text{-approx}} = qF(1-\alpha, p, q_{ref})/qF(1-\alpha, p, > q_{ref})\times 2$, and therefore $\text{RP}_{F\text{-approx}} > 2$.

The worked example in the next section describes a typical issue for field studies, of balancing the desire for more site replication to raise the regional scope of interest, against the raised error variation that accompanies sampling from more sites. For nested designs, different amounts of replication at the most nested level may involve no change of error d.f., in which case $\text{RP}_{F\text{-approx}} = n_{alt}/n_{ref}$. Consider the MM-SP design from the motivating example. With $r = 4$ replicate pots per combination of treatment levels, and $s = 3$ rooms per level of CO_2, its $n = 12$. This design would obtain twice the precision with $r = 8$ pots, giving $n = 24$ and therefore twice the relative performance ($\text{RP}_{F\text{-approx}} = qF(0.95, 1, 4)/qF(0.95, 1, 4)\times 24/12 = 2$, also by exact calculation). This doubling of n by using twice as many pots can therefore accommodate twice the error variance without loss of power. More than tripling the error variance would be accommodated, however, if the doubling of n is achieved instead by doubling the number of rooms per level of CO_2, from 3 to 6 (approximated by $qF(0.95, 1, 4)/qF(0.95, 1, 10)\times 24/12 = 3.11$).

3.4 Worked example

The barnacle *Semibalanus balanoides* has internal cross-fertilization in an entirely sessile adult stage. Reproduction therefore depends on living within penis-reach of neighbours. This life-history constraint inspired a field test of the hypothesis that larval settlement on rocky shores is promoted by local clusters of resident adults (Kent et al. 2003; Doncaster and Davey 2007). At the time of the study, the literature on the species was insufficient to identify a threshold response, below which any effect of cluster size could be deemed to cause negligible difference in reproductive success. Previous field trials had suggested a seemingly interesting effect of cluster size, however, with an estimated standardized effect size $\theta/\sigma = 1.6$ across replicate shores (from $\theta/\sigma = [(\text{MS[effect]} - \text{MS[error]})/(n\cdot\text{MS[error]})]^{0.5}$, as described in Kirk 1982).

The hypothesis was tested by measuring larval settlement density on patches of rock each cleared of resident barnacles except for a central cluster of a few cm^2 containing a set number of adults. The study aimed to draw conclusions relevant to shores with both high and low background levels of larval recruitment from pelagic waters. To this end, the cluster-size Treatment (B with 6 levels: 0, 2, 4, 8, 16, 32 adults per cluster, each replicated three times) was repeated across two replicate Shores (S') nested in Recruitment (A with 2 levels: low, high). This configuration sets the design as a replicated split plot for mixed model analysis: $\text{R}'_3(\text{B}_6|\text{S}'_2(\text{A}_2))$. Just two replicate shores would give power = 0.97 at $\alpha = 0.05$ to detect a main effect of treatment B even half the size of the previously estimated standardized effect size. The design stage nevertheless demanded an evaluation of the benefits of sampling from more replicate shores to achieve a wider geographical scope of inference, against the risk of thereby accruing enough random spatial variation to reduce power despite the raised error d.f.

Options for replication can be evaluated by measuring the performance of an alternative design with $s = 3$ shores per level of recruitment, relative to a reference with $s = 2$. The B and B × A terms of interest, both with 5 test d.f., share the same error term, B × S′(A) with $q = 10\times(s - 1)$ error d.f. Relative performance is thus approximated by the n-weighted ratio of critical F quantiles: $\mathrm{RP}_{F\text{-approx}} = qF(0.95, 5, 10)/qF(0.95, 5, 20)\times 18/12 = 1.84$. This means that sampling from one extra shore at each level of recruitment is expected to accommodate an 84 % increase in error variance without loss of power. This proxy for relative performance has further utility as a rule of thumb for the error fraction at any matching power. It indicates that power is likely to increase with more replication provided the higher n raises the error variance by less than 84 %, even when the prior estimate of power is based upon an imprecisely defined estimate of the standardized effect size.

The performance of the alternative relative to the reference design ($s = 3$ relative to $s = 2$) is enumerated exactly for a threshold power, of say 0.8 at $\alpha = 0.05$, by calculating the reference and alternative standardized effect sizes $\theta/\sigma = 0.602$ and 0.432 respectively for the B main effect (or 0.851 and 0.611 for B × A) required to achieve this power. Then the error fraction defining relative performance of the alternative design is precisely $\sigma^2_{alt}/\sigma^2_{ref} = (0.602/0.432)^2 = (0.851/0.611)^2 = 1.94$. The proxy value of 1.84 was therefore slightly conservative with respect to the true error fraction for matching a power of 0.8.

Kent et al. (2003) in fact attempted no such explicit assessments at the design stage. Adding more shores was deemed certain to increase the error variance, because it would require sampling along a wider stretch of coast that encompassed greater heterogeneity in geology and ocean currents. It was therefore decided by an implicit process to perform the study with just two shores per level of recruitment. The triple replication of treatment patches on each shore, however, anticipated the possibility of post-hoc pooling of error terms to raise the error d.f. (following Underwood 1997). In the event the B × S′(A) term was too close to significance to allow pooling, but main effects were anyway strongly significant and the B × A interaction far from significant. With the information now available on relative performance, the cost in raised error variance of adding two extra shores looks easily affordable, given the leeway of a near doubling in error variance provided by the benefit of higher q. A three-shore design could have used the same total of 72 patches with just two patch replicates per shore ($r\times b\times s\times a = 2\times 6\times 3\times 2$ instead of $3\times 6\times 2\times 2 = 72$), in which case it would have the same $n = 6$, and $\mathrm{RP}_{F\text{-approx}} = qF(0.95, 5, 10)/qF(0.95, 5, 20) = 1.23$.

4 Discussion

Studies that are designed for a specified effect threshold of interest can draw definitive conclusions following rejection or retention of the null hypothesis (Lenth 2001;

Bausell and Li 2002; Muller and Stewart 2006). The conclusions are definitive if the study is designed for acceptably low probabilities both of mistakenly rejecting the null hypothesis (the Type-I error at rate α), and of mistakenly retaining it (the Type-II error at rate β, from which power$= 1 - \beta$). However, much research in the biological sciences, and particularly in ecology and environmental sciences, concerns more preliminary stages of detecting treatment effects without presuming to know how large they must be to have an interesting influence on the system. Study designs for these stages are ill-suited to power analysis, because power depends on effect size. Here we have developed tools for evaluating design options at these stages for analyses of variance, where the choices involve different ways to structure nuisance variation and/or different numbers of replicates sampled at random from the population of interest.

Any study that tests only for the presence of real effects, without setting a minimum effect size of interest, cannot deliver definitive conclusions about non-significant treatments (Lipsey 1990; Hoenig and Heisey 2001; Colegrave and Ruxton 2003; Baguley 2004; Cumming 2008; Brosi and Biber 2009). Such studies must be regarded as provisional because of the possibility that non-significant treatments include real effects of potential interest that are too small in magnitude to be detected by the design. Provisional studies encompass not just the small-scale pilots to estimate parameters for power analysis, but all investigations with no predefined threshold of importance for the effect size. Amongst countless examples, consider a dataset of butterfly biodiversity that shows significant reduction in response to only some components of global climate change (e.g., Menendez et al. 2006). Any non-significant components cannot be dismissed until we know enough to define a threshold loss of negligible impact on ecosystems. Likewise, an experiment using *Daphnia* to test for effects of clonal diversity on competitive advantage cannot interpret non-significant effects without setting a threshold advantage of negligible impact on the community (Tagg et al. 2005).

Pilot or other studies can provide an estimate of the true error variance for prospective power analysis (Lenth 2001). They cannot inform on the true effect size, however, which is a desired output of the proposed study. For provisional studies, power analysis functions only to identify the ratio of true treatment effect size to error standard deviation that yields a target power for a putative design. Investigators then need to acknowledge the possibility that non-significant results may include real effects of smaller size than this. The estimation of relative performance circumvents this issue, and should therefore interest many experimental and field biologists. Since it holds the effect size and the power constant, the error variance on which power depends becomes an output, relative to a reference model, rather than an input requirement as for the estimation of relative efficiency. Although some knowledge of unmeasured variation is still required to interpret this output, it nevertheless informs on the work required of blocking, replication or other design features to absorb background variation.

The literature on statistical power focuses heavily on the problem of estimating power directly, usually in order to control its well-known sensitivity to error variance by optimising sample size (e.g., Lipsey 1990; Bausell and Li 2002; Faul et al. 2007; Maxwell et al. 2008). Comparisons of relative performance reverse this focus, by treating σ^2 as a response to power. As much as power is extremely sensitive to fractional changes in σ^2, so the error fraction is insensitive to power, as demonstrated in Figs. 2, 3 and 4. It makes sense to fix power in comparisons between alternative designs, since most prospective applications of power analysis aim to optimise designs for an acceptable level of power. Relative performance is a particularly useful concept at planning stages when little is known about the sizes either of treatment effects or variance components. For a given treatment, the choice between alternative designs for eliminating unmeasured or nuisance variation will be informed by evaluating how much needs to be absorbed in each alternative. Similarly, where the test question prescribes a particular design, a financial cost to replication can be calibrated against a benefit in greater scope of interest from accommodating proportionately more random variation. Conversely, a financial cost to controlling random variation can be calibrated against a proportionately cost-saving reduction in replication.

Our analysis of relative performance has throughout made the traditional ANOVA assumptions of equal variance, normality, and independence of errors (or conditional independence for a split plot analysis). Many ecological and environmental data violate one or more assumption. Classically, non-normality and heteroscedasticity were dealt with by transforming the response variable. However, transformation changes the relationship between effects in the model, e.g. additive effects on the mean become multiplicative effects on the median if the response is log transformed, and back transformation does not estimate the mean (Stanton and Thiede 2005).

If the data violate the normality assumption, a generalized linear model (Hardin and Hilbe 2012) may be appropriate. The methods proposed here are not needed for distributions with a fixed scale parameter, e.g. Poisson or Binomial, because then variance is a function of the mean and changing the experimental design will not necessarily change the variance. Relative performance will be useful for distributions with estimated scale parameters, e.g., the overdispersed Poisson or overdispersed Binomial distributions, when inference is based on a t or F statistic.

If the data violate the equal variances assumption, there are at least five approaches that could be used to make reasonable inferences: Welch's F test (Welch 1951), the modified F test (Brown and Forsythe 1974), White's heteroscedastic consistent variance estimator (White 1980), a transformation to homoscedasticity (Dutilleul and Potvin 1995; Dutilleul and Carrière 1998), and a Box-type adjustment (Brunner et al. 1997). The Welch's F test, modified F test, and Box-type adjustment change the computation of the F statistic and use a Satterthwaite approximation for the degrees of freedom. White's approach recalculates the variance of the parameter estimates, while Dutilleul's transformation preserves the group means while transforming the errors to independence and equal variance, for which the usual F test is appropriate. Each of these leads to an F statistic, so relative performance can be used to compare experimental designs. It will be harder to use relative performance with Welch's F test, the modified F test, or the Box-type adjustment because all three use a data-dependent estimated degrees of freedom, so the F quantiles used to calculate relative performance depend on more than the experimental design. Relative performance will be easiest to interpret when changing the experimental design has the same multiplicative effect on all variances.

If the data are correlated, they could be analysed using Dutilleul–Potvin–Carrière's transformation or a mixed model. Conditional on the model for the variance-covariance matrix of the errors, and conditional on the estimated parameters in that variance-covariance model, the $C\beta$ test statistics in a mixed model have F distributions (where C is a matrix of comparison coefficients expressing the null hypothesis, and β is the vector of parameters in a full rank or non-full rank model). Use of relative performance in the mixed model is complicated by data-dependent degrees of freedom, calculated either using a Satterthwaite approximation or a Kenward–Roger approximation (Kenward and Roger 1997). Relative performance will be easiest to interpret when the variance-covariance matrix can be written as $\sigma_1^2 V$ for one design and $\sigma_2^2 V$ for the other.

In conclusion, the n-weighted ratio of critical F quantiles provides a robust tool for exploring alternative design options for sustaining power. The only inputs needed are the critical values of F at given α and d.f., which are readily available from standard tables or programs, and the effective sample sizes, n. This proxy for relative performance has extra utility insofar as it approximates the error fraction at any matching power and therefore indicates that power is likely to increase in the event of the alternative design achieving a better error fraction. Although imprecise, the proxy fits the purpose of comparing designs for provisional studies, where power is estimated from best guesses of the treatment effects and the error variance, rather than being a defining component of the test question.

Acknowledgments This work was supported by grant NE/C003705/1 to CPD from the UK Natural Environment Research Council. Valuable comments and correctives were offered by K. E. Muller, R. W. Payne, W. J. Resetarits Jr, M. S. Ridout, and A. J. Vickers. Two anonymous reviewers made helpful suggestions for improving clarity.

Appendix: Derivation of *n*-weighted ratio of critical *F* quantiles for RP$_{F\text{-approx}}$

Given two designs with different error degrees of freedom, we wish to find the ratio of error variances that provides equal power for an arbitrary effect size.

Consider firstly the null hypothesis that the difference, δ, between two population means is zero. We can test $\delta = 0$ on two samples of size n with means$\bar{X}_1$ and $\bar{X}_2$ under the standard assumptions of independent, normally distributed errors with constant variance, v. This is done by computing a statistic $T = (\bar{X}_1 - \bar{X}_2)/\sqrt{2\,v/n}$, and comparing $|T|$ to $qt(1-\alpha/2, q)$, the $\alpha/2$ quantile of the t distribution with the appropriate q error d.f.

For a study with $\delta \neq 0$, the power of the T-test is estimated by assuming that the error variance, σ^2, equals v, and computing the non-centrality parameter, $\gamma = \delta/\sqrt{2 \cdot \sigma^2/n}$. Then $\beta = pt(qt(1-\alpha/2, q), q, \gamma)$, the cumulative frequency to $qt(1-\alpha/2, q)$ of a non-central t distribution with q d.f., and non-centrality parameter γ. Power $= 1 - \beta$.

Two different study designs for 1 test d.f. can be compared analytically by approximating the non-central t distribution by a γ-shifted t distribution (Anderson and Hauck 1983). That is, $pt(x, q, \gamma) \approx pt(x-\gamma, q)$. Setting $x = qt(1-\alpha/2, q)$, we obtain $\beta = pt(qt(1-\alpha/2, q), q, \gamma) \approx pt(qt(1-\alpha/2, q) - \gamma, q)$. At $\alpha = 0.05$, the approximate β lies within ±0.01 of all exact $\beta < 0.3$. The symmetrical t distribution with q d.f. has cumulative frequency $\beta = pt(qt(\beta, q), q)$, leading to the following approximation relating the difference in two means δ, the type-I error rate α, the type-II error rate β, the effective sample size n, and the error variance σ^2:

$$\delta = [qt\,(1-\alpha/2, q) - qt\,(\beta, q)] \cdot \sqrt{2 \cdot \sigma^2/n}. \tag{7}$$

We apply Eq. (7) to a reference design with effective sample size n_{ref}, error variance σ^2_{ref} and error d.f. q_{ref}, and to an alternative design with effective sample size n_{alt}, error variance σ^2_{alt} and error d.f. q_{alt}. If reference and alternative designs have the same power to detect the same difference δ, then

$$\begin{aligned}\delta &= [qt\,(1-\alpha/2, q_{alt}) - qt\,(\beta, q_{alt})] \cdot \sqrt{2 \cdot \sigma^2_{alt}/n_{alt}} \\ &= \left[qt\left(1-\alpha/2, q_{ref}\right) - qt\left(\beta, q_{ref}\right)\right] \cdot \sqrt{2 \cdot \sigma^2_{ref}/n_{ref}}\end{aligned} \tag{8}$$

Rearranging (8), the two designs have the same power when

$$\frac{\sigma^2_{alt}}{\sigma^2_{ref}} = \left[\frac{qt\left(1-\alpha/2, q_{ref}\right) - qt\left(\beta, q_{ref}\right)}{qt\,(1-\alpha/2, q_{alt}) - qt\,(\beta, q_{alt})}\right]^2 \times \frac{n_{alt}}{n_{ref}}. \tag{9}$$

Equation (9) approximates the error fraction required of an alternative design to match the power of a reference design, which we refer to as its relative performance (RP). For example, in a spatial treatment application, natural variability may be absorbed by using fewer replicates each of larger size (Sects. 3.3–3.4). A reference design with $n = 10$ replicates per sample has $q_{ref} = 18$, and an alternative design with $n = 5$ has $q_{alt} = 8$. The RP of the alternative design, in terms of its error fraction $\sigma^2_{alt}/\sigma^2_{ref}$ required to match a power of $1 - \beta = 0.8$ for the reference design, is approximated by:

$$\mathrm{RP}_{t\text{-approx}} = \left[\frac{qt\,(0.975, 18) - qt\,(0.2, 18)}{qt\,(0.975, 8) - qt\,(0.2, 8)}\right]^2 \times \frac{5}{10} = 0.430. \tag{10}$$

Thus the use of half as many replicates sustains 80 % power if it reduces natural variability by at least 57 %.

The t-test approach underlying Eqs. (9)–(10) applies only to tests of differences between 2 treatments or a 1-d.f. contrast among multiple treatments. We generalize relative performance to any F test with the approximation for $\sigma^2_{alt}/\sigma^2_{ref}$ at matching power:

$$\mathrm{RP}_{F\text{-approx}} = \frac{qF\left(1-\alpha, p, q_{ref}\right)}{qF\,(1-\alpha, p, q_{alt})} \times \frac{n_{alt}}{n_{ref}}, \tag{11}$$

where $qF(1-\alpha, p, q)$ is the α quantile of the F distribution with p numerator and q denominator d.f. associated with the test hypothesis. If the hypothesis has 1 d.f., then $qF(1-\alpha, 1, q) = qt(1-\alpha/2, q)^2$. In this case, $\mathrm{RP}_{F\text{-approx}}$ exactly equals $\mathrm{RP}_{t\text{-approx}}$ in four situations:

(1) When power = 0.5, so $qt(\beta, q) = 0$;
(2) When power = $1 - \alpha/2$, so $qt(\beta, q) = -qt(1-\alpha/2, q)$;

(3) When the ratios of quantiles are the same, so $qt(\beta, q_{ref})/qt(\beta, q_{alt}) = qt(1 - \alpha/2, q_{ref})/qt(1 - \alpha/2, q_{alt})$;
(4) When $q_{alt} = q_{ref}$, so $\text{RP} = n_{alt}/n_{ref}$; also by exact calculation.

For intermediate cases with 1 test d.f., $\text{RP}_{F\text{-approx}}$ is an approximation of $\text{RP}_{t\text{-approx}}$. For example, the Eq. (10) result of $\text{RP}_{t\text{-approx}} = 0.430$ has a corresponding $\text{RP}_{F\text{-approx}} = 0.415$ given by Eq. (11). The quality of this approximation is evaluated in Fig. 4a against precise calculation of $\sigma^2_{alt}/\sigma^2_{ref} = 0.428$ to sustain power = 0.8.

The application of Eq. (11) to $p > 1$ has a similar derivation. Here $\beta = pF(qF(1 - \alpha, p, q), p, q, \lambda)$, the cumulative frequency to the α quantile of a non-central F distribution with non-centrality parameter λ. This is approximated by a shifted F distribution (Patnaik 1949): $\beta \approx pF(qF(1 - \alpha, p, q)\cdot k, p', q)$, where $k = p/(p + \lambda)$ and $p' = (p + \lambda)^2/(p + 2\cdot\lambda)$. The F distribution with p' test d.f. and q error d.f. has $\beta = pF(qF(\beta, p', q), p', q)$, leading to the approximation:

$$k = qF(\beta, p', q)/qF(1 - \alpha, p, q). \tag{12}$$

Given the non-centrality parameter $\lambda = p\cdot n\cdot\theta^2/\sigma^2$, Eq. (12) rearranges to

$$\theta^2 = \left[\frac{qF\,(1 - \alpha, p, q)}{qF\,(\beta, p', q)} - 1\right] \times \frac{\sigma^2}{n}. \tag{13}$$

We apply Eq. (13) to a reference design with effective sample size n_{ref}, error variance σ^2_{ref}, test d.f. p and p'_{ref}, and error d.f. q_{ref}, and to an alternative design with effective sample size n_{alt}, error variance σ^2_{alt}, test d.f. p and p'_{alt}, and error d.f. q_{alt}. If reference and alternative designs have the same power to detect the same effect size θ, then the approximate error fraction of the alternative design is

$$\frac{\sigma^2_{alt}}{\sigma^2_{ref}} = \frac{qF\left(1 - \alpha, p, q_{ref}\right) - qF\left(\beta, p'_{ref}, q_{ref}\right)}{qF\left(1 - \alpha, p, q_{alt}\right) - qF\left(\beta, p'_{alt}, q_{alt}\right)} \times \frac{qF\left(\beta, p'_{alt}, q_{alt}\right)}{qF\left(\beta, p'_{ref}, q_{ref}\right)} \times \frac{n_{alt}}{n_{ref}}. \tag{14}$$

The values of the β quantiles of the F distribution in Eq. (14) are not precisely determinable for reference and alternative p' test d.f., which themselves depend on the unmeasured reference and alternative θ^2/σ^2. At small β, nevertheless, $\lambda_{alt} \approx \lambda_{ref}$ meaning that $p'_{alt} \approx p'_{ref}$, and the quantiles are relatively insensitive to q, such that $qF(\beta, p'_{alt}, q_{alt})/qF(\beta, p'_{ref}, q_{ref}) \approx 1$. In addition, low p gives these β quantiles values close to zero, resulting in Eq. (11) holding approximately for $\text{RP}_{F\text{-approx}}$. For $p > 10$, a closer approximation is

$$\text{RP}_{F\text{-approx}} = \frac{qF\left(1 - \alpha, p, q_{ref}\right) - 1}{qF\left(1 - \alpha, p, q_{alt}\right) - 1} \times \frac{n_{alt}}{n_{ref}}. \tag{15}$$

The quality of these approximations is evaluated in main-text Sect. 3 against precise calculations of $\sigma^2_{alt}/\sigma^2_{ref}$ to sustain power values of 0.99, 0.8, 0.5, 0.2 at $\alpha = 0.05$ for a range of designs and n.

References

Abou-el-Fittouh HA (1976) Relative efficiency of the randomized complete block design. Exp Agric 12:145–149

Abou-el-Fittouh HA (1978) Relative efficiency of the split-plot design. Exp Agric 14:65–72

Anderson S, Hauck WW (1983) A new procedure for testing equivalence in comparative bioavailability and other clinical trials. Commun Stat A-Theor 12:2663–2692

Bacchetti P (2010) Current sample size conventions: flaws, harms, and alternatives. BMC Med 8:17. http://www.biomedcentral.com/1741-7015/8/17

Baguley T (2004) Understanding statistical power in the context of applied research. Appl Ergon 35:73–80

Bausell RB, Li Y-F (2002) Power analysis for experimental research: a practical guide for the biological, medical and social sciences. Cambridge University Press, Cambridge

Blair RC, Higgins JJ, Karniski W, Kromrey JD (1994) A study of multivariate permutation tests which may replace Hotelling's T2 in prescribed circumstances. Multivar Behav Res 29:141–163

Brosi BJ, Biber EG (2009) Statistical inference, Type II error, and decision making under the US Endangered Species Act. Front Ecol Environ 7:487–494

Brown MB, Forsythe AB (1974) Small sample behaviour of some statistics which test equality of several means. Technometrics 16:129–132

Brunner E, Dette H, Munk A (1997) Box-type approximations in nonparametric factorial designs. J Am Stat Assoc 92:1494–1502

Cochran WG, Cox GM (1957) Experimental designs, 2nd edn. Wiley, New York

Colegrave N, Ruxton GD (2003) Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. Behav Ecol 14:446–450

Cumming G (2008) Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspect Psychol Sci 3:286–300

Doncaster CP, Davey AJH (2007) Analysis of variance and covariance: how to choose and construct models for the life sciences. Cambridge University Press, Cambridge. http://www.personal.soton.ac.uk/cpd/anovas/datasets/

Dutilleul P, Carrière Y (1998) Among-environment heteroscedasticity and the estimation and testing of genetic correlation. Heredity 80:403–413

Dutilleul P, Potvin C (1995) Among-environment heteroscedasticity and genetic autocorrelation: implications for the study of phenotypic plasticity. Genetics 139:1815–1829

Faul F, Erdfelder E, Lang A-G, Buchner A (2007) G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39:175–191

Fisher RA (1935, 1960) The design of experiments. Oliver and Boyd, Edinburgh

Hardin JW, Hilbe JM (2012) Generalized linear models and extensions, 3rd edn. Stata Press, College Station

Hinkelmann K, Kempthorne O (1994) Design and analysis of experiments, vol I. Wiley, New York

Hoenig JM, Heisey DM (2001) The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 55:19–24

Kent A, Hawkins SJ, Doncaster CP (2003) Population consequences of mutual attraction between settling and adult barnacles. J Anim Ecol 72:941–952

Kenward MG, Roger JH (1997) Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53:983–997

Kirk RE (1982) Experimental design: procedures for the behavioral sciences. Wadsworth, Belmont

Kraemer HC, Thiemann S (1987) How many subjects? Statistical power analysis in research. Sage, London

Legendre P, Dale MRT, Fortin MJ, Casgrain P, Gurevitch J (2004) Effects of spatial structures on the results of field experiments. Ecology 85:3202–3214

Lai K, Kelley K (2012) Accuracy in parameter estimation for ANCOVA and ANOVA contrasts: Sample size planning via narrow confidence intervals. Br J Math Stat Psychol 65:350–370

Lenth RV (2001) Some practical guidelines for effective sample size determination. Am Stat 55:187–193

Lenth RV (2006) Java applets for power and sample size [Computer software]. Retrieved August 3rd 2007, from http://www.stat.uiowa.edu/~rlenth/Power

Lipsey MW (1990) Design sensitivity: statistical power for experimental research. Sage, Newbury Park

Maxwell SE, Kelley K, Rausch JR (2008) Sample size planning for statistical power and accuracy in parameter estimation. Ann Rev Psychol 59:537–563

Menendez R, Megias AG, Hill JK, Braschler B, Willis SG, Collingham Y, Fox R, Roy DB, Thomas CD (2006) Species richness changes lag behind climate change. Proc R Soc Lond B 273:1465–1470

Muller KE, Stewart PW (2006) Linear model theory: univariate, multivariate, and mixed models. Wiley, New York

Neyman J, Iwaszkiewicz K, Kolodziejczyk St (1935) Statistical problems in agricultural experimentation. J R Stat Soc 2:107–180

Patnaik PB (1949) The non-central χ^2- and F-distributions and their applications. Biometrika 36:202–232

R Development Core Team (2010) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Rasch D, Herrendörfer G (1986) Experimental design: sample size determination and block designs. D Reidel, Dordrecht

Shieh G, Show-Li J (2004) The effectiveness of randomized complete block design. Stat Neerl 58:111–124

Stanton ML, Thiede DA (2005) Statistical convenience vs biological insight: consequences of data transformation for the analysis of fitness variation in heterogeneous environments. New Phytol 166:319–338

Steel RGD, Torrie JH (1960) Principles and procedures of statistics with special reference to the biological sciences. McGraw-Hill, New York

Tagg N, Innes DJ, Doncaster CP (2005) Outcomes of reciprocal invasions between genetically diverse and genetically uniform populations of *Daphnia obtusa* (Kurz). Oecologia 143:527–536

Underwood AJ (1997) Experiments in ecology: their logical design and interpretation using analysis of variance. Cambridge University Press, Cambridge

Verrill S, Durst M (2005) The decline and fall of Type II error rates. Am Stat 59:287–291

Vonesh EF (1983) Efficiency of repeated measures designs versus completely randomized designs based on multiple comparisons. Commun Stat A-Theor 12:289–301

Wang M, Hering F (2005) Efficiency of split-block designs versus split-plot designs for hypothesis testing. J Stat Plan Infer 132:163–182

Webb RY, Smith PJ, Firag A (2010) On the probability of improved accuracy with increased sample size. Am Stat 64:257–262

Welch BL (1951) On the comparison of several mean values: an alternative approach. Biometrika 38:330–336

White H (1980) A heteroscedastic-consistent covariance matrix estimator and a direct test for heteroscedasticity. Econometrika 48:817–838

Author Biographies

C. Patrick Doncaster is a Reader in Ecology at the University of Southampton whose research covers evolutionary ecology, population and community dynamics, and conservation. He is co-author with Andrew Davey of 'Analysis of variance and covariance: how to choose and construct models for the life sciences' published by Cambridge University Press in 2007.

Andrew J. H. Davey graduated with a PhD at Southampton in 2003, and went on to post-doctoral research at the National Institute of Water and Atmospheric Research in New Zealand. Since 2007 he has been at the UK-based WRc plc, where he is now their Senior Consultant statistician, specialising in the application of statistical techniques to environmental management problems.

Philip M. Dixon holds a professorial Chair in the Department of Statistics at Iowa State University. His research focuses on the development and evaluation of statistical methods to answer biological questions. These include methods for testing for negligible trend and for improving the precision of estimates of the frequency of rare events. He is on the research team of the Iowa State Climate Science Program.