Published February 19, 2014

Symposium: Statistical Concepts

Fundamentals of Experimental Design: Guidelines for Designing Successful Experiments

Michael D. Casler*

ABSTRACT

We often think of experimental designs as analogous to recipes in a cookbook. We look for something that we like, something that satisfies our needs, and frequently return to those that have become our long-standing favorites. We can easily become complacent, favoring the tried-and-true designs (or recipes) over those that contain unknown or untried ingredients or those that are too complex for our tastes and skills. Instead, I prefer to think of experimental designs as a creative series of decisions that are meant to solve one or more problems. These problems may be real or imagined—we may have direct evidence of a past or current problem or we may simply want insurance against future potential problems. The most significant manifestation of a "problem" or a "failed" design is unsatisfactory *P* values that prevent us from developing inferences about treatment differences. Four basic tenets or pillars of experimental design— replication, randomization, blocking, and size of experimental units— can be used creatively, intelligently, and consciously to solve both real and perceived problems in comparative experiments. Because research is expensive, both in terms of grant funds and the emotional costs invested in grant competition and administration, biological experiments should be designed under the mantra "failure is not an option." Guidelines and advice provided in this review are designed to reduce the probability of failure for researchers who are willing to question, evaluate, and possibly modify their decision-making processes.

Designing experiments involves a marriage of biological and mathematical sciences. The mathematical, or statistical, science is obvious. We use scientific fundamentals and principles that have been developed during the past century to conduct three types of experiments. Observational experiments are those designed to measure or verify an assumed constant, such as the velocity of light or the mass of an atom. Measurement experiments are those designed to measure the properties of a population, the members of which are variable, such as commodity prices, production statistics, or neutrino frequencies. As biologists, we are principally concerned with comparative (or manipulative) experiments, in which our global goal is to compare or contrast two or more practices or systems that may have some relevance to our field of scientific inquiry. It is the solutions, or the specific choices we must make, that are not so obvious.

In conducting comparative experiments, we routinely follow the general theory of scientific inquiry shown in Fig. 1. We begin with questions and/or hypotheses that must be translated to models based on the specific subject matter, e.g., cotton (*Gossypium hirsutum* L.) plants, farm machinery, or hay bales. The subject matter model is translated into a statistical model, which is developed in concert with the statistical design. The statistical design includes both the treatment design and the experimental design and provides a set of rules and procedures that allow us to conduct the experiment. For many of us, this process becomes routine, such that we tend to forget the fundamental nature and assumptions of statistical designs, instead forming designs and forging ahead using time-honored and traditional approaches that have worked well for us in the past. We often favor familiarity, simplicity, and constancy over any thought of change or concept of improvement. Researchers must recognize that designing comparative experiments is a massively decision-based exercise, so that the truism "If you choose not to decide, you still have made a choice" (Peart, 1980) is an appropriate concept in biological research.

Once the experiment has been conducted and the data collected, the statistical analysis proceeds as determined by the researcher before the experiment was conducted. The analysis leads to specific interpretations of the results, creating inferences and conclusions that bring the research back to the original question or hypothesis. Finally, and a point often ignored in most models, is the feedback loop that comes when the experiment is completed (Fig. 1). As scientists, we are intimately familiar with this feedback loop, often correctly opining that, "Any good experiment leads to more questions and new hypotheses than it answers." What we tend to forget or ignore is the fact that *two* types of information travel along this feedback loop: scientific answers to our biological questions and mathematical or statistical information that can be used to design better, more efficient, future experiments. Just as any experiment helps the

USDA-ARS, U.S. Dairy Forage Research Center, 1925 Linden Dr. W., Madison, WI 53706. Received 5 Mar. 2013. Accepted 4 May 2013. *Corresponding author (michaelcasler@ars.usda.gov).

Published in Agron. J. 107:692–705 (2015)
doi:10.2134/agronj2013.0114

Abbreviations: CRD, completely randomized design; RCBD, randomized complete block design.

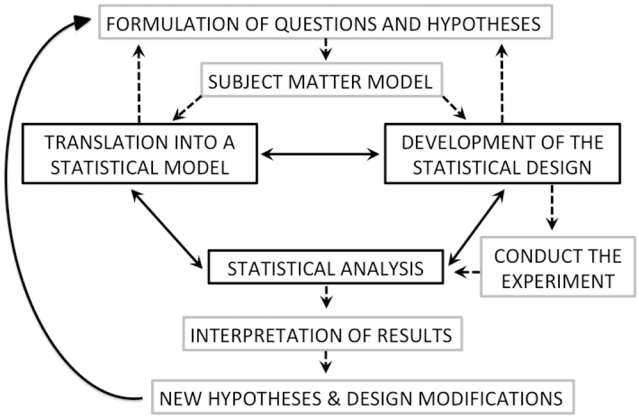


Fig. 1. Flow diagram of the logical steps in scientific experimentation, including a feedback loop that allows for new scientific hypotheses and experimental design modifications to future experiments. Boxes and arrows with heavier lines are directly related to the theme of this review. The three central boxes form the "statistical triangle" as described by Hinkelmann and Kempthorne (2008), the core of the statistical process.

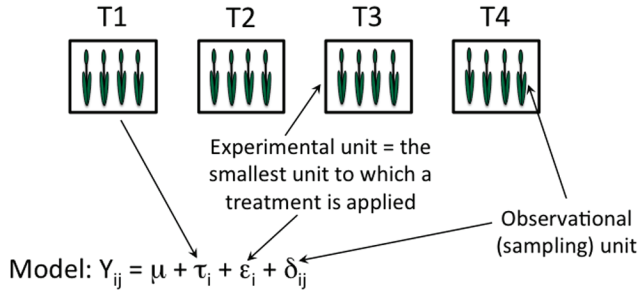
scientist to formulate new hypotheses, the completed experiment contains a wealth of information that can be used to help design better and more efficient experiments in the future. This is my principal focus here: how to design experiments with such a high probability of success that failure is exceedingly rare.

Of course, that begs the question, "What makes a better or more efficient experiment?" Before we can answer this question, we must first answer the question, "Is failure an option?" For the Mythbusters (Savage, 2009), "failure is always an option," because they are conducting measurement experiments for the purpose of answering simple questions with binomial answers, such as whether something occurs (or not) under specific circumstances. In their case, "failure" to prove a hypothesis is an acceptable result. For scientists, our research is so expensive, in both capital and emotional investment, that failure is distinctly NOT an option. Consider a simple experiment with two agronomic production systems. The goal is to compare the means of System A and System B and to develop an inference and set of conclusions from that result. What happens when the result of the experiment is a *P* value of 0.5? There are four potential reasons for this result: (i) a poorly designed experiment with insufficient power to detect a difference between the two means, (ii) poorly designed treatments that did not properly reflect the initial question or hypothesis, (iii) an improperly conducted experiment without proper oversight over treatment and data collection protocols, or (iv) lack of true differences between the treatment means. For my entire career, I have followed the philosophy of Frank N. Martin, former professor of statistics, University of Minnesota, "Everything is different from everything else." What Dr. Martin meant by this is simply, if a researcher formulates a valid question and puts sufficient thought into designing the treatments, then, by design, the treatments are different from each other. Failure to detect differences in treatment means is the fault of the experiment: a failure in the experimental design, the treatment design, the experimental conduct, the choice of measurement variables, or some combination thereof. Recognizing this, savvy referees and editors are frequently reluctant to accept manuscripts with "negative" results, especially when those results are based on an overwhelming lack of statistical significance. The underlying reason for lack of statistical significance or evidence of differences among treatment means can seldom, if ever, be resolved.

My purpose here is to review the literature and to provide guidelines and advice on how to avoid failure in comparative experimentation. Essentially, my purpose is to show readers how to develop personal and specific answers to the question, "What makes a better or more efficient experiment?" This review presents the concepts of experimental design as four pillars (replication, randomization, blocking, and experimental units), each of which must be given proper consideration and requires a conscious decision regarding number, size, scale, shape, or form. Each of these pillars has a profound impact on the experimental design, data analysis, and conclusions that result from the experimental conduct. As a prelude, Table 1 provides a list of terms that are widely used and essential to full comprehension of the contents of this review, including my personal definitions of each term, based both on my own experiences and on numerous textbook treatments, all of which are cited here.

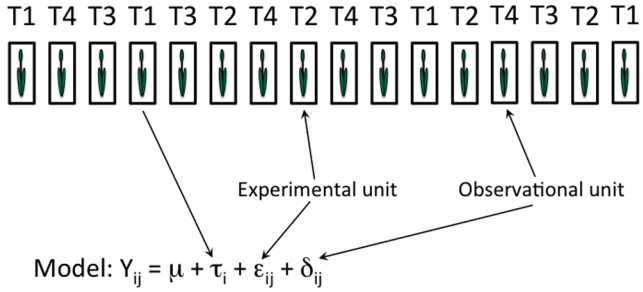
Table 1. Working definitions of statistical and experimental design terms used throughout this review.

Term	Definition
Experiment	a planned and organized inquiry designed to test a hypothesis, answer a question, or discover new facts
Treatment	a procedure or system whose effect on the experimental material is to be measured or observed
Experimental unit	the smallest unit to which a treatment is applied
Observational unit	the unit upon which observations or measurements are made
Block	a group of (presumably) homogeneous experimental units (a complete block contains all treatments)
Experimental design	the set of rules and procedures by which the treatments are assigned to experimental units
Treatment design	the organization or structure that exists across the treatments used to define the experiment
Replication	the practice of applying each treatment to multiple and mutually independent experimental units
Randomization	the practice of assigning treatments to experimental units such that each unit is equally likely to receive each treatment
Factor	a type of treatment; this can take on many forms: quantitative, qualitative, ranked, or nested
Level	a specific form or "state" of a factor
Factorial treatment	a combination of one level of each factor used to create a unique treatment combination
Experimental error	the variance among experimental units treated alike, often symbolized as σ^2 or σ_e^2.
Sampling error	the variance among observational units within experimental units; there can be multiple levels of sampling error
Precision	the inverse of experimental error, $1/\sigma_e^2$
Confounding	the purposeful or inadvertent mixing of two or more effects, such that no statistical analysis can separate them



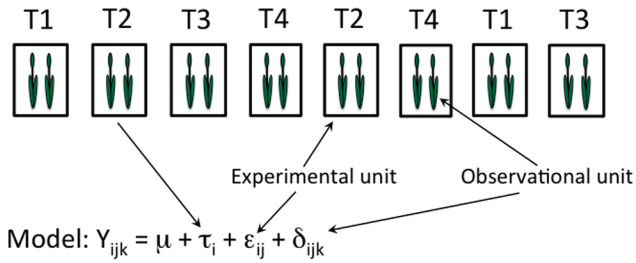
ANOVA Source of variation	df
Treatments + Experimental error	3 (fixed)
Observational error	12

Fig. 2. Design Example 1a: Sixteen plants are assigned to four experimental units, each of which is assigned to one treatment (T1–T4, symbolized by τ_i in the linear model). Treatments are confounded with experimental units because replication is conducted at a scale that does not match the treatment application.



ANOVA Source of variation	df
Treatments	3 (fixed)
Error (experimental + observational)	12

Fig. 3. Design Example 1b: Sixteen plants are assigned to 16 experimental units, four of which are independently assigned to each treatment (T1–T4, symbolized by τ_i in the linear model). This is an example of the completely randomized design.



ANOVA Source of variation	df
Treatments	3 (fixed)
Experimental error	4
Observational error	8

Fig. 4. Design Example 1c: Sixteen plants are assigned to eight experimental units, two of which are independently assigned to each treatment (T1–T4, symbolized by τ_i in the linear model). Each experimental unit contains two observational, or sampling, units. This is another example of the completely randomized design.

REPLICATION

Form and Scale of Replication

The act of replication serves four valuable functions in comparative experimentation. First, it provides a mechanism to estimate experimental error, which is essential to provide valid hypothesis tests and confidence intervals of estimators. Without replication, there is no ability to estimate background variation or "noise" in the estimates of treatment effects. Second, it provides a mechanism to increase the precision of an experiment. Based on the classic formula for a standard error, SE = $(\sigma^2/r)^{1/2}$, increasing the number of replicates has a direct, positive, and monotonic impact on experimental precision. Third, replication increases the scope of inference for the experiment. The act of replication draws a wider range of observations into the experiment, increasing the range of conditions encountered during the course of the experiment, especially if replication is conducted at multiple levels (explained below). Fourth, replication effects control of error. It puts the researcher in the driver's seat with regard to controlling the magnitude of experimental error and regulating the desired level of precision or power for the experiment.

While replication may occur at multiple levels or scales within an experiment, it must, first and foremost, be applied at the level of the experimental unit. Replicate observations must occur at a spatial and temporal scale that matches the application of the treatments (Quinn and Keough, 2002). Replication at the appropriate scale is essential because of the inherent variability that exists within biological systems and to avoid confounding treatment differences with other factors that may vary among experimental units. Consider an example in which 16 plants are arranged with four treatments; the treatments are applied, not to individual plants, but to groups of four plants (Fig. 2). Confounding treatments with experimental units creates two problems in the ANOVA or mixed models analysis: (i) the fixed effect of treatments contains an unknown random component associated with the unestimated experimental error; and (ii) the only estimable error term is the observational error, which is expected to be smaller than the experimental error. The net result is both an inflated *F* value and an inability to attribute "treatment" effects specifically to the treatments applied. A simple rearrangement, changing the experimental unit from a group of four plants to a single plant (Fig. 3), creates an acceptable scale of replication for treatments. In this case, observational and experimental errors are confounded, but the loss of information in that scenario is minor. Figure 4 represents a variation on this theme, in which there are also four observations made on each treatment, but the observations are organized into two sampling units within each of two independently assigned experimental units. The experimental units that receive Treatment 1 are independent of each other, but the sampling units (observational units) within each sampling unit are not independent of each other. Thus, there are two levels of replication: two replicates at the treatment level and two more within the treatment level.

Consider an experiment designed to test the effect of burning on the soil health of native grass prairies. Because fire is so difficult to manipulate on an experimental basis, the researcher chooses to conduct the experiment on two 1-ha plots or experimental units, one burned and one an unburned control. Numerous soil cores are taken from each plot at multiple points in time, creating both spatial and temporal levels of replication. An analysis of the data provides a valid hypothesis test of

a statistical difference between the two plots, both at a single point in time and over time points. However, the problem arises in interpretation of the results. If the null hypothesis is rejected, the next logical step is to interpret the results and discuss the reasons for the observed differences. Because the burning treatments were not replicated at the proper temporal or spatial scale, the comparison of burned vs. unburned is completely confounded with all other factors that differ between the two experimental units, including observed and unobserved factors. Even though the statistical analysis makes sense and is completely valid, the science is flawed—the experimental design does not allow an unequivocal interpretation of the results. Hurlbert (1984) coined the term *pseudoreplication* for this situation in which experimental treatments are replicated but at a scale that is not sufficient to provide a valid scientific interpretation of the results, i.e., complete confounding of treatments with experimental units (Fig. 2).

Two solutions exist to this problem: replication at the proper scale in either time or space. If the experiment is time critical, as is frequently the case with grant-funded research, replication in space may be the most desirable solution. Replication must be conducted at the level of the treatment: both burned and unburned treatments must be repeated across multiple experimental units, the experimental unit being a defined area that is burned or not burned. This could be done within a single prairie area or across multiple prairies, with prairies acting as blocks, each containing both treatments. If time is less critical and/or the experiment is limited to a small spatial area, then replication in time, e.g., across years, could be conducted at the proper scale but only if the treatments are independently randomized to experimental units in each year, i.e., the experiment is repeated on a new section of prairie in each year. In this manner, years would serve as both a replication and a blocking factor, allowing both a valid statistical hypothesis test and a reasonable level of scientific inference to be drawn.

One potential pitfall to the approach of using years or sites as a blocking factor arises if both the blocking factor and the treatment effect are considered fixed effects. Such a design is an application of the randomized complete block design (Cochran and Cox, 1957; Petersen, 1985; Steel et al., 1996) in which the block × treatment (B × T) interaction is generally used as the error term. Strictly speaking, an interaction between two fixed effects is also a fixed effect. If blocks, treatments, and their interaction are all fixed effects, there is no valid *F* test for treatments (Cochran and Cox, 1957; Quinn and Keough, 2002). Quinn and Keough (2002) advised testing for the presence of B × T interaction using Tukey's test for nonadditivity. Appendix 1 shows the SAS code for Tukey's single-degree-of-freedom test. If the null hypothesis of additivity is rejected, this suggests that either a transformation or an alternative distribution is warranted (Gbur et al., 2012). Alternatively, care should be taken in designing such experiments to allow the blocking factor to be treated as a random effect, e.g., in which each block could be considered a random observation from a larger population of levels. If the blocking factor can reasonably be assumed to have a random effect, then the B × T interaction can also be assumed to be a random effect, allowing it to serve as a valid error term.

When a treatment factor is not replicated at the proper scale, the confounding created by the lack of treatment replication extends through the entire experiment. For example, consider the experiment with two burning treatments described above and assume that each of the two plots contains 10 native grass species or mixtures replicated in a randomized complete block design. This design might appear to be a factorial treatment arrangement with a 2 × 10 treatment structure but in reality is two experiments, each conducted under different conditions (burned vs. unburned). Species are replicated at the proper scale but burning treatments are not. At the scale on which burning is applied, there are two treatments and two experimental units. While statistical software can always create a test statistic and *P* value for both main effects and the species × fire interaction, these tests are not necessarily valid or easily interpreted. The researcher eventually encounters difficulty in interpreting both the main effect of fire and the species × fire interaction. Because the main effect of fire is completely confounded with the two large experimental units, the biological or physical interpretation of the interaction suffers from the same fate. Due to the confounding, there is only a limited amount of interpretation that is possible due to the limitations of the experimental design. One cannot conclusively attribute a significant interaction to the effects of fire alone as a factor interacting with species. In a case such as this, analysis of the interaction effects should focus on the factor that is replicated at the proper scale, i.e., simple effects of species within fire treatments, using "fire" to represent different conditions under which the treatments are evaluated. In so doing, the larger factor, fire, is treated almost as a macroenvironmental factor, which forms the various conditions under which the main experiment was repeated. The prudent researcher would point out that "fire" is the defining variable but that other factors may be confounded with fire. This is analogous to the very common practice of repeating experiments across multiple locations and years for the purpose of assessing treatment differences under different environmental circumstances. As long as no attempt is made to attribute simple effects of interactions to unreplicated environmental variables, which cannot be conclusively resolved, the researcher is relatively safe from criticism.

The first "replication" step in designing most experiments is to explicitly define the experimental unit, the unit that forms the first level of replication (Table 1). Following establishment of the experimental unit and a decision about how to replicate treatments at the proper scale, additional levels of replication can be designed into the experiment, moving up the scale to larger units or down the scale to smaller units, or both. Two general questions are pertinent to this process: (i) what scales of replication are necessary to accomplish the experimental goals, and (ii) how much resources should be used to accomplish replication at each of the desired levels? The answer to the first question is largely biological but partly statistical. First and foremost, the scale of replication depends on the inferences desired and on exactly how data are to be collected. Replication at the scale of the experimental unit is almost always necessary to estimate a proper and unbiased error term. Experimental objectives that demand repetition across a range of environmental conditions suggest a need for additional replication at a scale larger than the experimental unit. Examples include locations, years, management factors, etc. At the other extreme, many variables cannot be measured on a whole experimental unit, demanding some

Table 2. Analysis of variance structure for an experiment conducted in a randomized complete block design repeated at multiple locations and years, assuming all effects to be random. Mixed models analysis provides direct estimates of the five variance components in the table, using restricted maximum likelihood estimation methods.

Source of variation†	df†	MS†	Expected values of mean squares†
Locations	$l - 1$		
Years	$y - 1$		
L × Y	$(l - 1)(y - 1)$		
Blocks within L × Y	$ly(r - 1)$		
Treatments	$t - 1$	MS_T	$\sigma^2 + r\sigma_{TLY}^2 + ry\sigma_{TL}^2 + rl\sigma_{TY}^2 + rly\sigma_T^2$
T × L	$(t - 1)(l - 1)$	MS_{TL}	$\sigma^2 + r\sigma_{TLY}^2 + ry\sigma_{TL}^2$
T × Y	$(t - 1)(y - 1)$	MS_{TY}	$\sigma^2 + r\sigma_{TLY}^2 + rl\sigma_{TY}^2$
T × L × Y	$(t - 1)(l - 1)(y - 1)$	MS_{TLY}	$\sigma^2 + r\sigma_{TLY}^2$
Experimental error	$ly(r - 1)(t - 1)$	MS_e	σ^2

† L, locations; Y, years; T, treatments, *l*, number of locations; *y*, number of years; *r*, number of blocks; *t*, number of treatments.

method of subsetting or subdividing experimental units into multiple observational units, i.e., replication at a scale below the experimental unit level. The answers to the second question lead to consideration of the number of replicates.

Number of Replicates

Two issues exist with respect to the number of replicates required to carry out an adequate experiment: (i) the number of replicates; and (ii) the distribution of replicates across the various forms of replication. A rich body of literature exists on the latter topic, generally falling under the topic of resource allocation (e.g., Brown and Glaz, 2001; Gordillo and Geiger, 2008; McCann et al., 2012).

Experimental replication can occur at four basic levels within the experiment: (i) the experimental unit, as discussed above; (ii) replication of the entire experiment, as with multiple locations and/or years; (iii) sampling at one or more levels within experimental units; or (iv) repeated measures. Classical resource allocation theory is based on having accurate estimates of the variance components from previous experiments. For example, consider an experimental situation in which field studies repeated at multiple locations are an annual activity, such as uniform cultivar evaluations (e.g., Yaklich et al., 2002). Decisions regarding the relative numbers of locations, years, and replicates are always critical in designing efficient future experiments. An ANOVA or mixed models analysis provides information that can be used to develop these inferences (Table 2), provided that the estimates have sufficient degrees of freedom to be accurate. Variance component estimates are inserted into the formula for the variance of a difference between two treatment means (V_{DTM}):

$$V_{DTM} = 2\left(\frac{\sigma_e^2}{rly} + \frac{\sigma_{TLY}^2}{ly} + \frac{\sigma_{TL}^2}{l} + \frac{\sigma_{TY}^2}{y}\right)$$

where *r*, *l*, and *y* are the numbers of replicates, locations, and years of future potential experiments, respectively, and the variances are defined in Table 2. It is always clear from these exercises that the number of locations and/or years should be maximized relative to the number of replicates (e.g., McCann et al., 2012). The only situation that favors maximizing the number of replicates is when all the treatment × environment interactions are zero, which is rarely the case. Having recognized this general principle for many years, numerous plant breeders conduct family-based selection protocols using multiple locations, chosen to represent the target population of environments, with only a single replicate of each family present at each location. While such a design is often considered heretical for publication purposes, it can be extremely resource efficient for the purposes of maximizing efficiency in a breeding program. This methodology can be extended to nearly any form of replication, both larger than the experimental unit and smaller than the experimental unit. The general guideline to replicate the largest units to the maximum allowed by the experimental budget and logistical restrictions is nearly always the rule.

The number of replicates and the distribution of those numbers across the various forms of replication is one of the most difficult and complex decisions in designing an experiment. Based on my experiences teaching and consulting on this topic, this is the single most important and most often ignored topic in experimental design. The number of replicates has a direct, highly predictable, repeatable, and tangible effect on precision and the ability to detect differences among treatments. Despite this fact, few agronomic researchers have ever conducted a meaningful power analysis with the goal of designing a better experiment.

Power is the probability of rejecting a null hypothesis that is, in fact, false. Following Dr. Martin's philosophy, properly designed treatments, evaluated or measured with the proper measurement variables and properly administered throughout the experiment, will result in significant differences between treatment means *only* when a sufficient level of replication was used. If all else was done properly, the failure to detect differences among treatment means was due to an insufficient and/or improper scale of replication. Designing experiments that are expected to have a high level of power is our principal mechanism to avoid this result.

Gbur et al. (2012, Ch. 7) provided the most extensive and thorough treatment available for developing power analyses, providing numerous examples and SAS code, including variables that follow non-normal distributions and different levels of replication. The following steps are required to implement the Gbur et al. (2012) probability distribution method to estimate the power of a future hypothetical experiment:

1. Obtain an estimate of experimental error variability from previous experiments conducted in the appropriate field or laboratory setting or from the literature.
2. Identify the distribution of the variable of interest, e.g., normal, binomial, Poisson, binary, etc.
3. Determine the *P* value that will be used to set detection limits [$\alpha = P$(Type I error)] and the minimum difference between treatments to be detected (δ).
4. Choose an experimental design structure that includes all desired blocking arrangements and one or more levels of replication.

5. Create an exemplary data set that matches the desired experimental design structure and contains two "dummy" treatments with constant values across all experimental and observational units of the data set. To create such an exemplary data set, a single value must be chosen for all the replication factors in the hypothetical experiment.

Appendix 2 shows the SAS code to accomplish a single power analysis in three coding steps:

1. Create an exemplary data set for a completely randomized design with two treatments, four experimental units per treatment (replicates), and four observational units per experimental unit.
2. Compute the noncentrality parameter of the F distribution with the appropriate degrees of freedom under the alternative hypothesis that the two treatments have difference δ.
3. Compute the power from the noncentrality parameter.

Appendix 3 shows a modified version of this SAS code that contains a macro, automating the code to allow estimation of the power for numerous experimental design parameterizations. This code estimates the power for this design with a number of replicates ranging from four to eight and the number of observations per experimental unit ranging from two to four, with a wider range of values displayed in Fig. 5. Such a graphical display easily allows any researcher to choose one of multiple design arrangements that meet the expected target for power, which is often set at 0.8 (80%). For example, the results of this exercise easily validate the statements made above regarding resource allocation, that the largest sized replication factors (experimental units in this case) have the greatest direct impact on improving future experiments. In the case of Fig. 5, using five different values for the number of observational units per experimental unit, any one of five different scenarios can be chosen for use in future experiments, each with an expectation of having power $\approx$0.8.

Exercises such as this one can be created for any experimental design and sampling scheme imaginable (Gbur et al., 2012). Furthermore, the effectiveness of power predictions can be evaluated directly in terms of empirical detection limits for completed experiments. For example, Casler (1998) completed an extensive and expensive series of experiments that were designed to provide means and variance estimates to select for increased forage yield in several species and using several breeding methods. The next step, following selection and synthesis of the progeny populations, was to compare the progeny populations created by the various selection methods. Because very small differences between populations were expected, and the goal was to detect these differences with a reasonable P value, a power analysis was essential before planning the second experiment (Appendix 4). While statistical theory, common sense, and the power analysis all suggested the need for four to six locations (Fig. 6), practical considerations led to the choice of three locations and 16 replicates in a randomized complete block design (Casler, 2008). The net result of this power exercise was a series of experiments with LSD values ranging from 2 to 3% of the mean and a high frequency of P values <0.01 (Casler, 2008), i.e., an extremely successful and satisfying result.

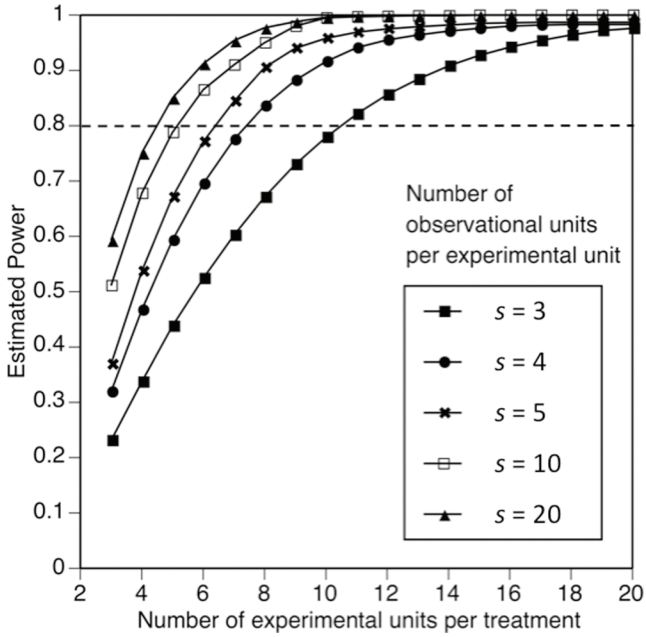


Fig. 5. Estimated power of a hypothesis test designed to detect a treatment difference of 5% of the mean with a Type I error rate of 0.05, variance component estimates of 5 and 10 (experimental and sampling errors, respectively), and varying numbers of experimental units and observational units (s = 3–20). The dashed line represents power = 0.8 and illustrates that different replication and sampling scenarios can be created to achieve the same result.

Unreplicated Experiments

Numerous special situations exist for which there is a strong temptation or need to devote all resources toward multiple treatments and none to replication or independent observations of those treatments. Many on-farm experiments represent a classic example of this situation. Farmers and outreach personnel see value in trying different treatments on a very large scale but see little value in replication of those treatments. Researchers organizing these types of field experiments have three options: (i) conduct the experiments on multiple farms, using farms as blocks; (ii) use control-plot designs in which one control treatment is interspersed with the other treatments, optimally in an every-other-plot pattern; or (iii) use a combination of both approaches. Control-plot designs were developed in the early 20th century (Pritchard, 1916; Richey, 1924) and their popularity lasted through the 1970s, largely for use in evaluating extremely large numbers of lines or families for plant breeding (Baker and McKenzie, 1967; Mak et al., 1978). These designs fell out of favor when it became clear that devoting half of the experimental units to a single treatment or cultivar was highly inefficient (Melton and Finckner, 1967), and modern spatial analysis methods were just emerging (Gaeton and Guyon, 2010). Nevertheless, they remain a very viable option for small on-farm experiments.

Augmented designs represent a specific form of design that can handle hundreds or thousands of treatments, most of which are unreplicated. Treatments are generally cultivars or breeding lines that must be evaluated over multiple locations and for which seed supplies are often highly limiting and extremely valuable (Casler et al., 2000, 2001). Augmented designs involve a large number of unreplicated treatments organized into very small blocks, often blocks of only five

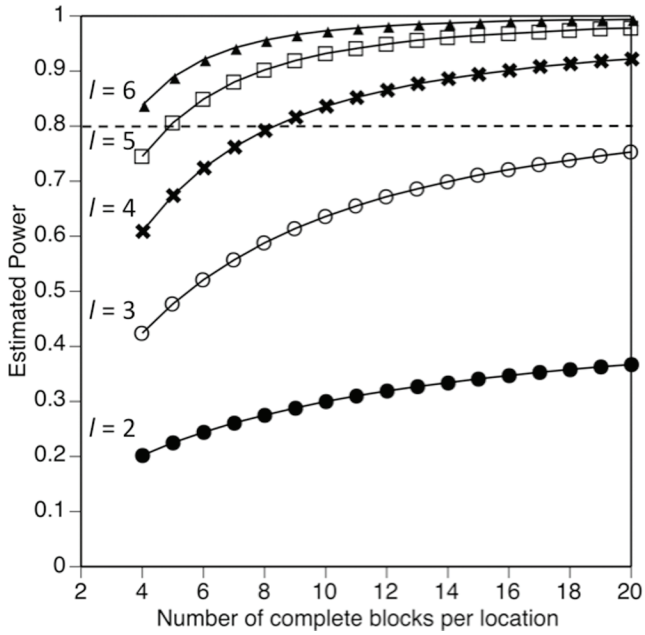


Fig. 6. Estimated power of a hypothesis test designed to detect a treatment difference of 10% of the mean with a Type I error rate of 0.05, variance component estimates of 0.02 and 0.2 (treatment × location interaction and experimental error, respectively), and varying numbers of locations (l = 2–6) and blocks. The dashed line represents power = 0.8 and illustrates that different combinations of number of locations and replicates can be created to achieve the same or similar expected results.

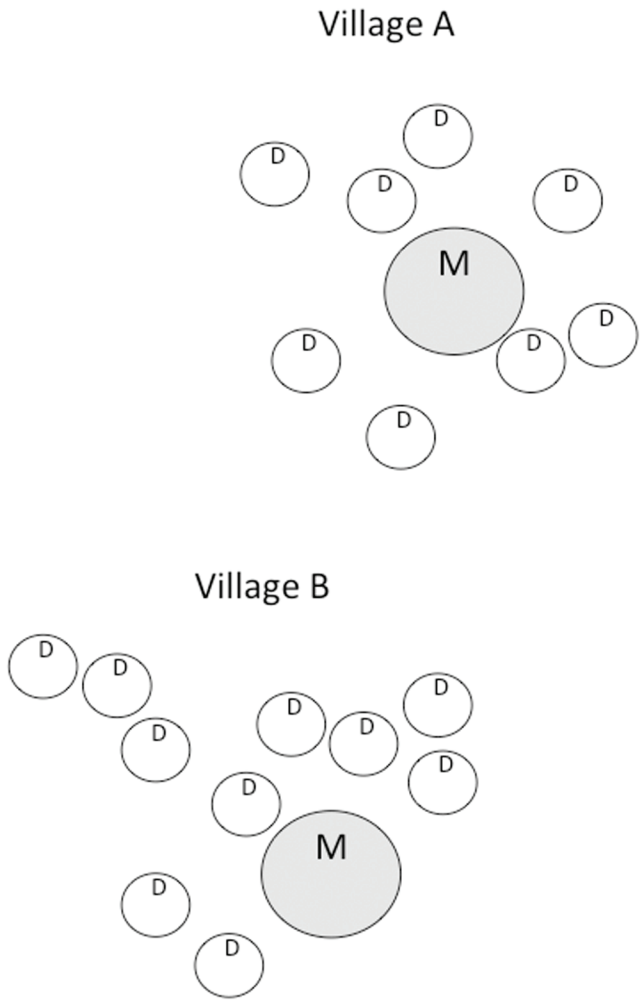


Fig. 7. Illustration of mother–daughter experimental designs for two villages of participatory researchers, in which mother trials (M) are conducted by researchers on experimental stations and daughter trials (D) are conducted by individual farmers on their landholdings.

experimental units (Lin and Poushinsky, 1983, 1985; Casler et al., 2000, 2001). The center plot within each block consists of a single cultivar that serves the purpose of error estimation and spatial adjustment on a large scale. Additional check cultivars are added at random to individual blocks to augment error estimation and to estimate the spatial effects on a finer or smaller scale (Lin and Poushinsky, 1983, 1985).

A variation of this theme has recently emerged in the field of participatory research: mother–daughter designs (Morrone and Snapp, 2001; Snapp et al., 2002; van Eeuwijk et al., 2001). These designs are generally focused on a small number of practical treatments in which farmers or small landholders have a genuine interest and/or a sense of ownership. The most traditional application involves the use of a "mother" trial, conducted with proper replication and randomization on an experimental research station, accompanied by numerous "daughter" trials conducted by various landholders (Fig. 7). Mother trials generally include all possible treatments of interest in a region or village, while daughter trials include only a small number of treatments of interest to the participant or landholder. Each participant is free to design and/or choose the treatments of greatest interest and value for the individual farm. Daughter trials are linked to each other and to the mother trial by the fact that all treatments are represented in the mother trial and some treatments are repeated across multiple daughter trials (van Eeuwijk et al., 2001).

Lastly, situations occasionally arise in which experimental units are so expensive that the tradeoff between replication and additional treatments may favor the latter. As an example, to be of greatest value to both the practical and scientific communities, grazing research should be conducted at multiple levels or stocking rates (animals per hectare). Because grazing research facilities are limited in the number of discreet paddocks available for use as independent experimental units, and because experimental units are very costly, it is generally impossible to include all three desired factors: treatments, grazing levels, and replicates. Bransby (1989) proposed unreplicated designs, with regression-based statistical analysis of multiple treatments repeated across multiple grazing levels, as an alternative to classically replicated designs.

RANDOMIZATION

The principle of randomness applies to the proper conduct of experiments at two levels. First, a careful definition of the experimental materials and facilities to be included in the experiment demands that each be properly sampled to ensure that it is properly represented. Whether the treatments derive from bags of seed, livestock, pathogens, or field sites meant to represent one or more environmental variables, the population must be defined and a random or representative sample must be chosen to represent the population. Populations can be defined very broadly with a clear intent to choose a random sample. This is often the case when the desired inference is to represent a population that is larger than the sample, leading to the choice of a random effect (Fig. 8). Conversely, when an inference is desired

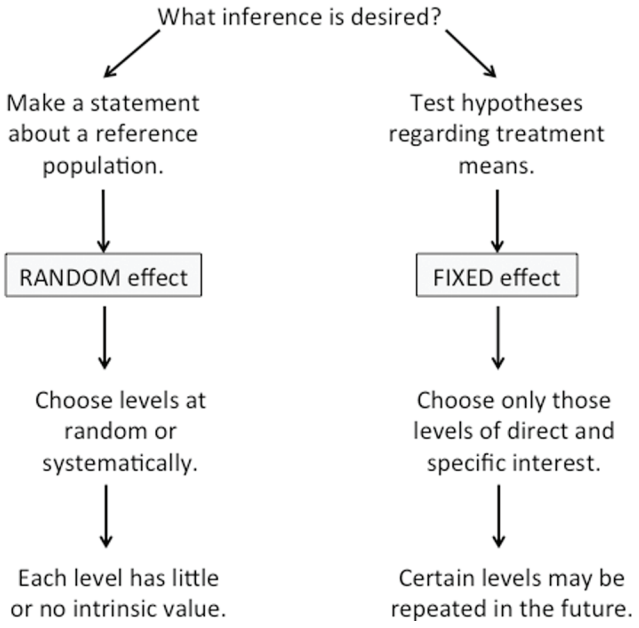


Fig. 8. Flow diagram illustrating the decision rule for fixed vs. random effects for each factor in an experiment. The first step is to ask the question about the desired inference, which leads to the decision of fixed vs. random. Following this decision, the experimental treatments can be chosen as desired to meet the chosen inference.

only for a small number of levels, each of which can be included in the experiment, the choice is usually to treat this as a fixed effect. Selecting a random sample remains a critical component of defining and applying the treatments, even for fixed effects, to ensure that the inference matches the hypothesis.

The second aspect of randomization concerns the assignment of treatments to experimental units. By definition, in a properly randomized experiment, every treatment is equally likely to be applied to every experimental unit. The simplest application of this definition is to randomly apply r replicates of t treatments to rt experimental units, without regard to order, structure, or priority. This is the completely randomized design (CRD), which consists of one block, represented in both Fig. 3 and 4. More complicated designs, involving hierarchical or multistep randomization are discussed below.

Randomization provides two key services in experimental designs: (i) unbiased estimation of treatment means and experimental errors; and (ii) a precaution against a disturbance that may or may not arise. Randomization is our insurance policy; like insurance, we must pay for it, not with funds, but with some time spent conducting randomizations and some inconvenience in the conduct of the experiment; however, the benefits nearly always outweigh the costs. If there were no field gradients, no hidden spatial variation, and independent observations of our treatments provided uniform results, we would not need to randomize or replicate our treatments; we would be chemists or physicists. Of course, the field of biology is rife with hidden sources of variability, including many sources that are unknown or unexpected during the course of conducting an experiment. Randomization helps to ensure that those unknown or unexpected sources of variation do not introduce biases, confounding, or elimination of valid hypothesis tests during the course of the experiment.

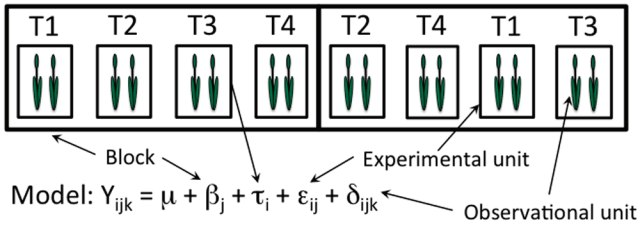
This brings up the issue of which is better: randomization or interspersion (Hurlbert, 1984). Randomization is a strict mathematical process in which a random number generator is used to order the treatments. Interspersion is a subjective concept in which the researcher strives to achieve balance and avoid "clumping" of certain treatments or replicates of treatments. Sometimes randomization can result in experimental designs that appear to have undesirable clumping or patterns. These patterns can result in potential bias or confounding to treatment effects if there are underlying spatial variation patterns that are correlated with the randomization plan (Martin, 1986). They can also lead to significant differences in the level of precision for statistical comparisons: treatments that have high average distances between replicates will tend to have higher variances than those with low average distances between replicates (van Es and van Es, 1993). Spatially balanced complete block designs (van Es et al., 2007) were developed to solve this potential problem, helping to ensure that both randomization and interspersion are key components of any design. These designs promote spatial balance among treatment comparisons and do not allow treatments to occur in the same position in different blocks.

BLOCKING

Blocking is utilized in experimental designs for one or both of two purposes: (i) for precision, to create groups of experimental units that are more homogeneous than would occur with random sampling of the entire population of experimental units; or (ii) for convenience, to allow different sizes of experimental units when larger plots or larger experimental areas are required for the application of one factor compared with other factors. In addition, much like randomization, blocking can be thought of as an insurance policy against disturbances that may or may not arise during the course of the experiment. I can cite examples of the use of randomized complete block designs (RCBD) being used to successfully recover from post-establishment disturbances to colleagues' experiments, including military maneuvers (three different incidents), a house mover, wayward herbicide sprayers (twice), a motorcycle gang, an angry and vengeful former employee, and partying students.

The RCBD design is the simplest blocking design (Fig. 9), and it can be used with any of the treatment structures shown in Table 3. The only restriction on the RCBD design is that block size is equal to the number of treatments. Because some factorial experiments can become very large very quickly, composite designs and fractional factorials represent two treatment designs that are intended to reduce the number of treatments from a full factorial to a meaningful subset designed to focus on specific treatment comparisons. Reducing the number of treatments in this manner reduces both the cost of the experiment and the block size, potentially improving the precision for hypothesis tests.

Treatment design is frequently a driver of experimental design (Table 4). Indeed, many experimental designs exist only as a randomization restriction of the basic RCBD design. The most common of these is the split-plot family of designs, which exist only as a specific randomization restriction that is placed on certain factorial experiments (Fig. 10). A common misconception of the split-plot randomization is that they represent an experimental design per se, e.g., a split-plot with Factor A as the whole-plot factor and Factor B as the subplot factor. In fact, this represents an incomplete description of the experimental



ANOVA Source of variation	df
Blocks	1
Treatments	3 (fixed)
Experimental error	3
Observational error	8

Fig. 9. Design Example 2: Sixteen plants are assigned to eight experimental units. Experimental units are grouped into two blocks, each of which contains a number of experimental units exactly equal to the number of treatments (T1–T4, symbolized by τ_i in the linear model). Each experimental unit contains two observational, or sampling, units. This is an example of the randomized complete block design with sampling.

design because it is missing the basic design for Factor A, the whole-plot factor. For example, the design in Fig. 10 illustrates a RCBD with two blocks for Factor A, with Factor B acting as a "split" of each Factor A whole plot. More precisely, Fig. 10 represents a split-plot randomization restriction of a RCBD design. As an illustration, Casler et al. (2000, 2001) used a split-plot randomization of a Latin square design in which the whole-plot factor was arranged in a 5 × 5 or 6 × 6 Latin square.

Split-plot randomizations are extremely flexible and versatile, applicable to factorial treatment designs with two or more factors. They are frequently used strictly for convenience, when the whole-plot factor requires larger experimental units than the subplot factor(s). Common examples include tillage treatments, irrigation treatments, and planting dates. Their versatility is illustrated by the fact that multiple splits can be easily incorporated into the design, either for the purpose of logistics and convenience or for statistical purposes. A further variation is the split block or strip plot, in which there are two whole-plot factors that are stripped across each other. Care must be taken to ensure that both whole-plot factors are randomized independently and differently in each replicate of the experiment, rather than stripping one factor across the entire experiment without rerandomization. Combined use of both strip-plot and traditional split-plot "splits" in one experiment (Riesterer et al., 2000) illustrate both the complexity and versatility available in these randomization restrictions.

In the simplest split-plot design, with two factors, the statistical concept is illustrated as follows. There are two error terms, one for the whole plot (E_a) and one for the subplot (E_b). Their expected values are: $E(E_a) = \sigma^2(1 + \rho)$ and $E(E_b) = \sigma^2(1 - \rho)$, where σ^2 is the unit variance and ρ is the autocorrelation coefficient. The statistical success of the design relies on the empirical relationship $E_a > E_b$, which is caused by values of $\rho > 0$. Split plots that are used for statistical reasons are meant to take advantage of this relationship, such that precision is increased for the subplot factor and the interaction at the expense of precision on the whole-plot factor and whole-plot-based simple effects. Researchers who design experiments using the split-plot concept for convenience will often hope for $\rho \approx 0$, or $E_a \approx E_b$, especially when they prefer not to have a difference in precision between the whole-plot and subplot factors.

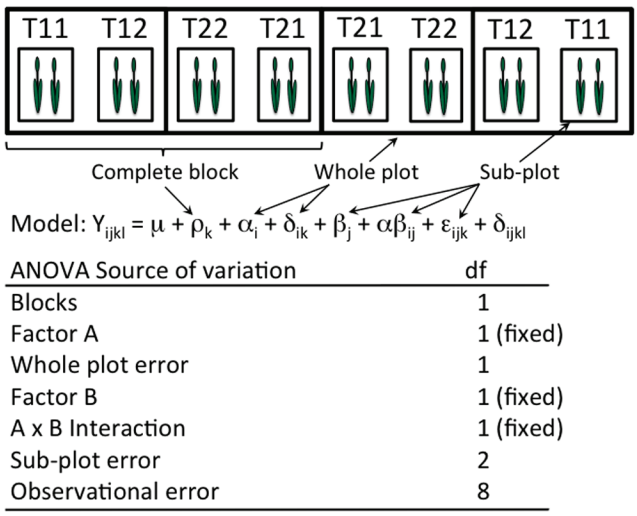
Table 3. List of some common treatment structures utilized in designing comparative or manipulative experiments.

Number of factors	Design name or definition	Defining characteristics
One	unstructured	there is no structure or organization to the treatments
Two or more	nested structure	factors have levels that are not repeated or have the same meaning at all levels of the other factors (Schutz and Cockerham, 1966)
Two or more	full factorial design	each factor has a specific number of levels that are repeated (have the same definition and meaning) over all levels of the other factors; the number of treatment combinations is the product of the number of levels of each factor (Cochran and Cox, 1957)
Two or more	confounding design	a full factorial in which a higher order interaction term is sacrificed as a blocking factor in order to achieve a reduction in block size (Cochran and Cox, 1957; Cox, 1958)
Two or more	composite design	a subset of a factorial designed to severely reduce the number of treatments required to evaluate the main effects and first-order interactions using regression-based modeling (Draper and John, 1988; Draper and Lin, 1990; Lucas, 1976)
Two or more	fractional factorial	a partial factorial arrangement in which only a subset of the factorial treatments is included, usually based on choosing a higher order interaction term as the defining contrast (Cochran and Cox, 1957; Cox, 1958)
Two or more	repeated measures	one or more of the treatment factors is observed over multiple time points without rerandomization of treatments to experimental units (Milliken and Johnson, 2009)

Blocking designs that involve arranging the blocks in a linear manner, without any prior knowledge of spatial variation among the experimental units, can severely reduce the probability of success for an experiment. Because many experimental research stations consist of large and visually uniform fields, bidirectional blocking is an excellent insurance policy against guessing wrong on the blocking arrangement. Numerous experimental design families are available for purposeful bidirectional blocking (Table 4). The simplest of these is the Latin square, in which the number of treatments is restricted to equal the number of replicates. Due to this restriction, Latin square designs are uncommon in field research, although they can be highly effective for small experiments (Casler et al., 2007). Incomplete block designs, such as the lattice square and the incomplete Latin square, relax this requirement and offer considerable options for both reducing block size and creating bidirectional blocking to remove spatial variability in two directions. Alpha designs and row–column designs offer additional flexibility and options for varying the number of treatments, number of replicates, and block size (John and Eccleston, 1986; Patterson and Williams, 1976; Williams, 1986). In particular, these designs have become very common in tree breeding

Table 4. Experimental design families organized according to type and complexity of blocking arrangements.

Number of potential blocking levels	Treatment design (from Table 3)	Experimental design options	Defining characteristics	References
One	any structure	randomized complete block	block size = number of treatments (t)	Steel et al. (1996)
Multiple and flexible	full factorial	split plot and variations, split block (strip plot)	design contains multiple sizes of experimental units, one for each level (or "split"); larger experimental units may be required for convenience but will be associated with increased error if Error(a) > Error(b)	Cochran and Cox (1957), Cox (1958), Petersen (1985), Steel et al. (1996)
Bidirectional and structured	any structure	Latin square, Graeco-Latin square, lattice square, incomplete Latin square	bidirectional blocking in perpendicular directions; number of replicates and treatments are highly restricted in some designs	Cochran and Cox (1957), Petersen (1985), Steel et al. (1996)
Multiple and flexible, bidirectional	any structure	alpha, row–column	treatments arranged in rows and columns; extremely flexible with regard to number of treatments, number of replicates, and block size	John and Eccleston (1986), Patterson and Robinson (1989), Williams (1986)
Multiple and flexible	nested structure	blocks in reps (sets in reps), reps in blocks (reps in sets)	treatments are randomly divided into sets; block size = number of treatments per set; good for inferences on random effects	Schutz and Cockerham (1966), Casler (1998)
Multiple and flexible	any structure	balanced or partially balanced incomplete blocks	potentially large reduction in block size with flexibility in both structure and field layout; block size (k) may be $t^{1/2}$ or $t^{1/3}$	Cochran and Cox (1957), Cox (1958), Petersen (1985)
Multiple and flexible	any structure with variable replication across treatments	control plot, augmented, modified augmented	designed to accommodate extremely unbalanced treatment structures and unequal replication across treatments; no restrictions on numbers of treatments and replicates	Chandra (1994), Lin and Poushinsky (1983, 1985), Wolfinger et al. (1997)



ANOVA Source of variation	df
Blocks	1
Factor A	1 (fixed)
Whole plot error	1
Factor B	1 (fixed)
A x B Interaction	1 (fixed)
Sub-plot error	2
Observational error	8

Fig. 10. Design Example 3: Sixteen plants are assigned to eight experimental units. Experimental units are grouped into two blocks, each of which contains a number of experimental units exactly equal to the number of treatments (T11–T22). Within each block, the experimental units are blocked again according to two levels of Factor A (α_i). Each whole unit or whole plot of Factor A is further subdivided according to two levels of Factor B (β_j). This is an example of a randomized complete block design with a split-plot randomization restriction (Factor A = whole plots; Factor B = subplots).

programs that are typically located in mountainous regions with little or no level or flat topography.

The last three families of designs in Table 4 each contain considerable flexibility and versatility intended to solve particular problems. The blocks-in-reps and the reps-in-blocks designs were designed for large random-effects experiments in plant and animal breeding, with the goals of precisely estimating genetic parameters of populations (Schutz and Cockerham, 1966). These designs continue to be useful in breeding and selection experiments. Balanced and partially balanced incomplete block designs represent a diverse family of designs meant to create large numbers of very small incomplete blocks that allow all potentially important treatment comparisons to be made. Cochran and Cox (1957, Ch. 9, 10, and 11) provided a thorough treatment of these designs, complete with numerous plans and analytical details. Lastly, control-plot designs (Baker and McKenzie, 1967; Mak et al., 1978), augmented designs (Murthy et al., 1994; Scott and Milliken, 1993; Wolfinger et al., 1997), and modified augmented designs (Lin and Poushinsky, 1983, 1985; Lin and Voldeng, 1989) were designed to accommodate large numbers of treatments in which there is considerable lack of uniformity in the number of replicates per treatment or some treatments are completely unreplicated.

Finally, there is a direct, positive correlation between design complexity and efficiency. This relationship is moderated by experiment size and the nature of spatial variability. In general, the simplest designs (CRD and RCBD) are the least efficient designs, an effect that is magnified as the number of treatments increases or as spatial variability becomes stronger and more unpredictable. Either a large number of treatments or unpredictable spatial variation can severely undermine the efficiency of the CRD and RDBD. The RCBD is extremely popular due to its simplicity (van Es et al., 2007), but when the number of treatments is >20, all researchers should consider the general principle that more blocks and smaller blocks are nearly always more efficient than complete blocks (Legendre et al., 2004; Lin and Voldeng, 1989). Numerous options exist for some relatively simple incomplete block designs that allow efficient estimation and control of both predictable and unpredictable spatial variation and create favorable interspersion patterns for treatment randomizations.

SIZE OF EXPERIMENTAL UNITS

The last pillar of experimental design is the least understood and possesses the least amount of theoretical results to support empirical observations. This subject receives little or no coverage in the textbooks that deal with experimental design (e.g., less

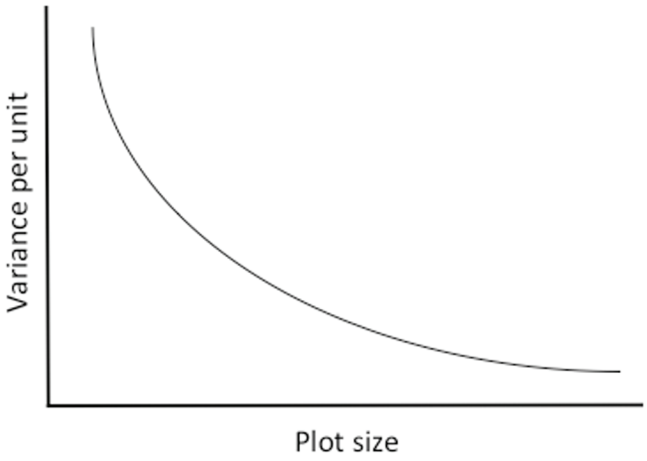


Fig. 11. Graphical illustration of Smith's Law (Smith, 1938), demonstrating the empirical relationship between variance per unit and plot size in agronomic field experiments.

than one page in Steel et al., 1996), perhaps owing to the lack of theoretical results. The concept of optimal plot (experimental unit) size is very old, predating Fisher's early concepts of experimental design and analysis of variance. The concept is based on Smith's Law (Smith, 1938), which derived from the general observation of a negative asymptotic relationship between variance (on a per-unit or single-plot basis) and plot size (Fig. 11). Smith's Law seems to be a general phenomenon that holds across a wide range of species and environmental conditions (LeClerg, 1966). Smith's Law is most frequently modeled with log-linear regression equations, the slope of which (b) may vary widely across species and environments (LeClerg, 1966), reflecting differential levels of spatial homogeneity.

The asymptotic nature of this relationship creates a conundrum for any researcher who seeks to modify plot size and to predict the statistical consequences of such a change (Fig. 11). For small initial plot sizes, any change in plot size is expected to have a large effect on mean unit variance. Conversely, for researchers who are using relatively large plots, relatively large changes in plot size may have little or no impact on the mean unit variance. Of course, the distinction between "small" and "large" plots is completely relative, probably varying with species, soil types, measurement variables, and numerous other factors. In my own research program, we consider a "small" plot to be the smallest area that can be planted with standard equipment (five-row drilled plots, 0.9 by 1.2 m). Conversely, we consider a "large" plot to be the longest distance that can be planted with one rotation of the cone planter without having significant gaps in seed placement (15 m) and a width equal to any multiple of 0.9 m (multiple passes of the cone planter). For any given experiment, our choice of plot size depends on seed and land availability and prior estimates of the slope of the log-linear curve in Fig. 11, estimated for that particular field, as described below.

The question for everyone who is considering either an increase in plot size to achieve a decrease in variance, or a decrease in plot size to reduce the experimental area, is, "where are my experiments located on this curve?" Lin and Binns (1984, 1986) used Smith's Law to develop an empirical method for estimating the impact of changes in plot size on experimental error variances. They developed a series of simple post-analysis computations from any blocking design, allowing researchers to compare numerous experimental designs for future experiments based on varying plot size, number of treatments (block size), number of replicates (blocks), and total experiment size. This method is focused on empirical estimation of b, Smith's coefficient of soil heterogeneity, leading to some generalizations. If $b < 0.2$, increase the number of replicates or blocks, utilizing incomplete block designs if the number of treatments is large. Reductions in plot size can be advantageous but only if they allow an increase in the number of replicates. If $b > 0.7$, increasing the plot size is likely to be more cost effective than increasing the number of replicates. If $0.2 < b < 0.7$, increasing either or both plot size and number of replicates should be effective.

Because these inferences are derived from statistics estimated from ANOVA or mixed models output, they represent random variables that are subject to errors of estimation, such as sampling variation and environmental effects. Their greatest utility will probably derive from many years of experimentation at a site maintaining historical records of estimates of b, along with critical data such as field number, orientation within the field, and basic experimental design characteristics, such as the number of replicates, number of treatments, block size, and plot size.

SUMMARY AND CONCLUSIONS

Biological research is expensive, with monetary costs to granting agencies and emotional costs to researchers. As such, biological researchers should always follow the mantra, "failure is not an option." A failed experimental design is generally manifested as an experiment with high P values, leaving the researcher with uncertain or equivocal conclusions: are the treatments really not different from each other, is my experimental design faulty due to poor planning and decision making, or was there some unknown and unseen disturbance that occurred to the experiment, causing errors to be inflated? Rarely can these questions be answered when P values are high, resulting in unpublishable results and wasted time and money. To borrow an experimental design term, these causal explanations are confounded with each other when treatment effects are nonsignificant. It is generally impossible to assign cause to one or the other explanation.

Researchers who have the benefit of long-term research results, accumulated from many years of research at a particular site or on a defined group of subjects, have access to a wealth of data that can be used for planning purposes. Such a database can be used to derive accurate estimates of variances and expectations for treatment mean differences, empowering the researcher to conduct meaningful power analyses. Resource allocation exercises, combined with the estimation of heterogeneity coefficients, can guide decisions as to the optimum distribution of various forms of replication, plot size, block size (number of treatments per block), and block orientation. Many of the computations that are required to develop these inferences are fairly simple and routine, making it simple and easy to maintain a spreadsheet or database that summarizes the experimental design implications from all experiments with common goals and measurement variables.

For young researchers who do not have access to such a database, the challenge of increasing the probability of success for experimental designs is formidable. While field sites may appear uniform, experience tells us that they are probably nonuniform

and that patterns of spatial variability cannot be predicted without years of trial and error. My advice to researchers in this situation is to follow a few basic guidelines in your early years:

- Become reasonably proficient in experimental design and data analysis or become good friends with someone who has this qualification and can benefit from working with you.
- Begin by designing small experiments with relatively large experimental units, as large as you feel you can handle with labor and machinery.
- Always conduct a thorough power analysis before designing every experiment, even if you have to make guesses about parameter estimates.
- If you must design large experiments, use incomplete block designs that are simple to set up and analyze but sufficiently flexible to generate small blocks that can be arranged to account for bidirectional gradients in the field.
- Begin maintaining the database suggested above.
- Do not become complacent. Push yourself beyond your current comfort zone. Stretch the limits of your imagination, knowledge base, and experiences.
- "Come on you raver, you seer of visions, come on you painter, you piper, you prisoner, and shine!" (Waters et al., 1975).

APPENDIX 1

SAS code to compute Tukey's test for nonadditivity in a randomized complete block design. The test is computed in two parts: (1) output the predicted values from the mixed models analysis, then (2) square the predicted values and include this term as a single-degree-of-freedom regressor variable (covariate) as a test for multiplicative block and treatment effects (Sahai and Ageel, 2000).

```
data a; infile 'tukey.dat';
input rep block trt y;
proc mixed; class trt;
model pcb = trt / outpred=x;
random block;
data x; set x;
p2=pred*pred;
proc mixed; class trt;
model pcb = trt p2;
random block;
```

APPENDIX 2

This SAS code to estimate the power of a hypothesis test has the following parameters: treatment means = 95 and 100, variance components = 5 and 10 (experimental error and sampling error, respectively), $r = 4$ replicates, $s = 2$ observational units per experimental unit, and the assumption of normally distributed errors (adapted from Gbur et al., 2012).

```
options nocenter mprint;
data a; input trt y;
do rep=1 to 4 by 1;
do samples=1 to 2 by 1;
output;
end;
end;
datalines;
1 95
2 100
run;
proc glimmix; class trt rep;
model y = trt;
random rep(trt);
parms (5)(10) / hold=1,2;
ods output tests3=power_terms;
data power;
set power_terms;
alpha=0.05;
ncparm=numdf*Fvalue;
F_critical=finv(1–alpha, numdf, dendf, 0);
power=1–probf(F_critical, numdf, dendf, ncparm);
proc print;
run;
```

APPENDIX 3

The SAS code from Appendix 1 is embedded in a macro that allows power to be estimated for a range of experimental designs with four to eight experimental units per treatment (rep, repl, repmax) and two to four observational units per experimental unit (obs, obsv, obsmax).

```
options nocenter mprint;
%macro one(obsmax,repmax);
data a;
%do obsv=2 %to &obsmax;
group1=&obsv;
%do repl=2 %to &repmax;
group2=&repl;
do obs=2 to &obsv by 1;
do rep=4 to &repl by 1;
do trt=0 to 1 by 1;
output;
end;
end;
end;
%end;
%end;
%mend one;
%one(4,8);      /*  change values here  */
run;
proc sort; by group1 group2;
data b; set a; by group1 group2;
if trt=0 then y=95;
if trt=1 then y=100;
run;
proc glimmix; class trt rep; by group1 group2;
model y = trt;
random rep(trt);
parms (5)(10) / hold=1,2;
ods output tests3=power_terms;
```

```
data power;
set power_terms;
alpha=0.05;
ncparm=numdf*Fvalue;
F_critical=finv(1–alpha, numdf, dendf, 0);
power=1–probf(F_critical, numdf, dendf, ncparm);
proc print;
run;
```

APPENDIX 4

This SAS code to estimate the power of a hypothesis test has the following parameters: treatment means = 9 and 10, variance components = 0.02 and 0.2 (treatment × location interaction and experimental error, respectively), $r = 4$ to 20 replicates, $l = 2$ to 6 locations, and the assumption of normally distributed errors (adapted from Gbur et al., 2012).

```
options nocenter mprint;
%macro two(locmax,repmax);
data a;
%do locn=2 %to &locmax;
group1=&locn;
%do repl=4 %to &repmax;
group2=&repl;
do loc=2 to &locn by 1;
do rep=4 to &repl by 1;
do trt=0 to 1 by 1;
output;
end;
end;
end;
%end;
%end;
%mend two;
%two(6,20);        /*  change here   */
run;
proc sort; by group1 group2;
data b; set a; by group1 group2;
if trt=0 then y=9;
if trt=1 then y=10;
run;
proc glimmix; class location trt rep; by group1 group2;
model y = trt;
random trt*location;
parms (0.02)(0.2) / hold=1,2;
ods output tests3=power_terms;
data power;
set power_terms;
alpha=0.05;
ncparm=numdf*Fvalue;
F_critical=finv(1–alpha, numdf, dendf, 0);
power=1–probf(F_critical, numdf, dendf, ncparm);
proc print;
run;
```

REFERENCES

Baker, R.J., and R.I.H. McKenzie. 1967. Use of control plots in yield trials. Crop Sci. 7:335–337. doi:10.2135/cropsci1967.0011183X000700040017x

Bransby, D.I. 1989. Compromises in the design and conduct of grazing experiments. In: G.C. Marten, editor, Grazing research: Design, methodology, and analysis. CSSA Spec. Publ. 16. ASA and CSSA, Madison, WI. p. 53–67.

Brown, J.S., and B. Glaz. 2001. Analysis of resource allocation in final stage sugarcane clonal selection. Crop Sci. 41:57–62. doi:10.2135/cropsci2001.41157x

Casler, M.D. 1998. Genetic variation within eight populations of perennial forage grasses. Plant Breed. 117:243–249. doi:10.1111/j.1439-0523.1998.tb01933.x

Casler, M.D. 2008. Among-and-within-family selection in eight forage grass populations. Crop Sci. 48:434–442. doi:10.2135/cropsci2007.05.0267

Casler, M.D., S.L. Fales, A.R. McElroy, M.H. Hall, L.D. Hoffman, and K.T. Leath. 2000. Genetic progress from 40 years of orchardgrass breeding in North America measured under hay management. Crop Sci. 40:1019–1025. doi:10.2135/cropsci2000.4041019x

Casler, M.D., S.L. Fales, D.J. Undersander, and A.R. McElroy. 2001. Genetic progress from 40 Years of orchardgrass breeding in North America measured under management intensive rotational grazing. Can. J. Plant Sci. 81:713–721. doi:10.4141/P01-032

Casler, M.D., K.P. Vogel, C.M. Taliaferro, N.J. Ehlke, J.D. Berdahl, E.C. Brummer, R.L. Kallenbach, C.P. West, and R.B. Mitchell. 2007. Latitudinal and longitudinal adaptation of switchgrass populations. Crop Sci. 47:2249–2260. doi:10.2135/cropsci2006.12.0780

Chandra, S. 1994. Efficiency of check-plot designs in unreplicated field trials. Theor. Appl. Genet. 88:618–620. doi:10.1007/BF01240927

Cochran, W.G., and G.M. Cox. 1957. Experimental designs. John Wiley & Sons, New York.

Cox, D.R. 1958. Planning of experiments. John Wiley & Sons, New York.

Draper, N.R., and J.A. John. 1988. Response-surface designs for quantitative and qualitative variables. Technometrics 30:423–428. doi:10.1080/00401706.1988.10488437

Draper, N.R., and D.K.J. Lin. 1990. Small response-surface designs. Technometrics 32:187–194. doi:10.1080/00401706.1990.10484634

Gaeton, C., and X. Guyon. 2010. Spatial statistics and modeling. Springer, New York.

Gbur, E.E., W.W. Stroup, K.S. McCarter, S. Durham, L.J. Young, M. Christman, M. West, and M. Kramer. 2012. Analysis of generalized linear mixed models in the agricultural and natural resources sciences. ASA, CSSA, and SSSA, Madison, WI.

Gordillo, G.A., and H.H. Geiger. 2008. Alternative recurrent selection strategies using doubled haploid lines in hybrid maize breeding. Crop Sci. 48:911–922. doi:10.2135/cropsci2007.04.0223

Hinkelmann, K., and O. Kempthorne. 2008. Design and analysis of experiments. Vol. 1. Introduction to experimental design. 2nd ed. Wiley-Interscience, New York.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54:187–211. doi:10.2307/1942661

John, J.A., and J.A. Eccleston. 1986. Row-column a-designs. Biometrika 73:301–306.

LeClerg, E.L. 1966. Significance of experimental design in plant breeding. In: K.J. Frey, editor, Plant breeding. Iowa State Univ. Press, Ames. p. 243–313.

Legendre, P., M.R.T. Dale, M. Fortin, P. Casgrain, and J. Gurevitch. 2004. Effects of spatial structures on the results of field experiments. Ecology 85:3202–3214. doi:10.1890/03-0677

Lin, C.S., and M.R. Binns. 1984. Working rules for determining the plot size and numbers of plots per block in field experiments. J. Agric. Sci. 103:11–15. doi:10.1017/S0021859600043276

Lin, C.S., and M.R. Binns. 1986. Relative efficiency of two randomized block designs having different plot sizes and numbers of replications and plots per block. Agron. J. 78:531–534. doi:10.2134/agronj1986.00021962007800030029x

Lin, C.S., and G. Poushinsky. 1983. A modified augmented design for an early stage of plant selection involving a large number of test lines without replication. Biometrics 39:553–561. doi:10.2307/2531083

Lin, C.S., and G. Poushinsky. 1985. A modified augmented design (Type 2) for rectangular plots. Can. J. Plant Sci. 65:743–749. doi:10.4141/cjps85-094

Lin, C.S., and H.D. Voldeng. 1989. Efficiency of Type 2 modified augmented designs in soybean variety trials. Agron. J. 81:512–517. doi:10.2134/agronj1989.00021962008100030024x

Lucas, J.M. 1976. Which response surface design is best. Technometrics 18:411–417. doi:10.1080/00401706.1976.10489472

Mak, C., B.L. Harvey, and J.D. Berdahl. 1978. An evaluation of control plots and moving means for error control in barley nurseries. Crop Sci. 18:870–873. doi:10.2135/cropsci1978.0011183X001800050049x

Martin, R.J. 1986. On the design of experiments under spatial correlation. Biometrika 73:247–277. doi:10.1093/biomet/73.2.247

McCann, L.C., P.C. Bethke, M.D. Casler, and P.W. Simon. 2012. Allocation of experimental resources to minimize the variance of genotype mean chip color and tuber composition. Crop Sci. 52:1475–1481.

Melton, B., and N.D. Finckner. 1967. Relative efficiency of experimental designs with systematic control plots for alfalfa yield tests. Crop Sci. 7:305–307. doi:10.2135/cropsci1967.0011183X000700040006x

Milliken, G.A., and D.E. Johnson. 2009. Analysis of messy data. Vol. 1. Designed experiments. 2nd ed. CRC Press, Boca Raton, FL.

Morrone, V.L., and S.S. Snapp. 2001. Uptake of a new on-farm trial design that includes the small-holder farmer. HortScience 36:477.

Murthy, N.S., R.V.S. Rao, and C.R. Nageswara Rao. 1994. Screening of tobacco germplasm lines using modified augmented design. Tob. Res. 20:91–96.

Patterson, H.D., and D.L. Robinson. 1989. Row-and-column designs with two replicates. J. Agric. Sci. 112:73–77. doi:10.1017/S0021859600084124

Patterson, H.D., and E.R. Williams. 1976. A new class of resolvable incomplete block designs. Biometrika 63:83–92. doi:10.1093/biomet/63.1.83

Peart, N. 1980. Freewill. Rush, Permanent Waves, Island/Mercury Records Ltd., London.

Petersen, R.G. 1985. Design and analysis of experiments. Marcel Dekker, New York.

Pritchard, F.J. 1916. The use of check plots and repeated plantings in varietal trials. J. Am. Soc. Agron. 8:65–81. doi:10.2134/agronj1916.00021962000800020001x

Quinn, G.P., and M.J. Keough. 2002. Experimental design and data analysis for biologists. Cambridge Univ. Press, Cambridge, UK.

Richey, F.D. 1924. Adjusting yields to their regression on a moving average, as a means of correcting for the heterogeneity. J. Agric. Res. 27:79–90.

Riesterer, J.L., M.D. Casler, D.J. Undersander, and D.K. Combs. 2000. Seasonal yield distribution of cool-season grasses following winter defoliation. Agron. J. 92:974–980. doi:10.2134/agronj2000.925974x

Sahai, H., and M.I. Ageel. 2000. Analysis of variance: Fixed, random, and mixed models. Springer, New York.

Savage, A. 2009. Mythbusters. Episode 114: Demolition derby. The Discovery Channel. http://dsc.discovery.com/tv-shows/mythbusters

Schutz, W.M., and C.C. Cockerham. 1966. The effect of field blocking on gain from selection. Biometrics 22:843–863. doi:10.2307/2528078

Scott, R.A., and G.A. Milliken. 1993. A SAS program for analyzing augmented randomized complete block designs. Crop Sci. 33:865–867. doi:10.2135/cropsci1993.0011183X003300040046x

Smith, H.F. 1938. An empirical law describing heterogeneity in the yields of agricultural crops. J. Agric. Sci. 28:1–23. doi:10.1017/S0021859600050516

Snapp, S., G. Kanyama-Phiri, B. Kamanga, R. Gilbert, and K. Wellard. 2002. Farmer and researcher partnerships in Malawi: Developing soil fertility technologies for the near-term and far-term. Exp. Agric. 38:411–431. doi:10.1017/S0014479702000443

Steel, R.G.D., J.H. Torrie, and D.A. Dickey. 1996. Principles and procedures in statistics. 3rd ed. McGraw-Hill, New York.

van Eeuwijk, F.A., M. Cooper, I.H. DeLacy, S. Ceccarelli, and S. Grando. 2001. Some vocabulary and grammar for the analysis of multi-environment trials, as applied to the analysis of FPB and PPB trials. Euphytica 122:477–490. doi:10.1023/A:1017591407285

van Es, H.M., and C.L. van Es. 1993. Spatial nature of randomization and its effect on the outcome of field experiments. Agron. J. 85:420–428. doi:10.2134/agronj1993.00021962008500020046x

van Es, H.M., C.P. Gomes, M. Sellman, and C.L. van Es. 2007. Spatially-balanced complete block designs for field experiments. Geoderma 140:346–352. doi:10.1016/j.geoderma.2007.04.017

Waters, R., R. Wright, and D. Gilmour. 1975. Shine on you crazy diamond. Pink Floyd, Wish You Were Here, Columbia/CBS Records, New York.

Williams, E.R. 1986. Row and column designs with contiguous replicates. Aust. J. Stat. 28:154–163. doi:10.1111/j.1467-842X.1986.tb00594.x

Wolfinger, R.D., W.T. Federer, and O. Cordero-Brana. 1997. Recovering information in augmented designs, using SAS PROC GLM and PROC MIXED. Agron. J. 89:856–859. doi:10.2134/agronj1997.00021962008900060002x

Yaklich, R.W., B. Vinyard, M. Camp, and S. Douglass. 2002. Analysis of seed protein and oil from soybean northern and southern region uniform tests. Crop Sci. 42:1504–1515. doi:10.2135/cropsci2002.1504