CBCS
HTTP://AGROBIOL.SGGW.WAW.PL/CBCS

COMMUNICATIONS IN BIOMETRY AND CROP SCIENCE
VOL. 8, NO. 1, 2013, PP. 23–38

INTERNATIONAL JOURNAL OF THE FACULTY OF AGRICULTURE AND BIOLOGY,
WARSAW UNIVERSITY OF LIFE SCIENCES, POLAND

REGULAR ARTICLE

Finding Hidden Treasure: A 28-Year Case Study for Optimizing Experimental Designs

Michael D. Casler

USDA-ARS, 1925 Linden Dr., Madison, WI 53706-1108 USA.
E-mail: michael.casler@ars.usda.gov

CITATION: Casler, M.D. (2013). Finding Hidden Treasure: A 28-Year Case Study for Optimizing Experimental Designs. *Communications in Biometry and Crop Science* 8 (1), 23–38.

Received: 14 September 2012, Accepted: 13 August 2013, Published online: 18 September 2013

ABSTRACT

Field-based agronomic and genetic research is a decision-based process. Many decisions are required to design, conduct, analyze, and complete any field experiment. While these decisions are critical to the success of any research program, their importance is magnified for research on perennial crops due to multiple years of data collection. The objective of this paper is to summarize 28 years of field-based perennial forage grass research at a single location describing changes to experimental design methodology, illustrating both predicted and empirical results of those changes. The study is based on an analysis of total forage yield for 114 genetic experiments of 11 forage grass species. Over the course of time, plot sizes were reduced from 5.6 to 2.8 to 1.4 m^2, resulting in a decrease in mean CV from 18.6 to 13.3 to 11.5%, respectively. These changes in precision, directly opposite that predicted from Smith's Law of Heterogeneity, were attributed largely to a vastly improved relative efficiency of blocking and spatial adjustment as plot size was decreased: 212 vs. 130% relative efficiency of blocking and 240 vs. 109% relative efficiency of spatial adjustment for 1.4 vs. 5.6- m^2 plots. These changes suggested that spatial variation at this site consists of fine-scale variation that is uneven, unpredictable, and cannot be easily captured by incomplete blocking or spatial analyses of the larger experimental units. Finally, a power analysis was used to predict the number of replicates required to detect expected differences for a series of experiments, resulting in a high level of predictability and a highly successful application of power analysis to assist with the design of field experiments.

Key Words: *blocking; coefficient of variation; spatial analysis; precision; experimental design.*

INTRODUCTION

Field-based agronomic and genetic research is a decision-based process. Many decisions are required to design and conduct a field experiment, collect and analyze the data, and interpret the results. A large number of these decisions have nothing to do with the hypotheses to be tested, but instead relate to the design of the experimental arrangement used to create valid and convenient hypothesis tests. Common decisions include size and shape of the experimental unit, number of replicates, randomization restrictions, block size and shape, method of data analysis, type and extent of replication (related to the desired range of inferences), and duration of the experiment.

Most of these decisions are made using one or both of two broad criteria. First, many elementary statistical textbooks and a few journal articles offer some general guidelines on broad concepts of field-plot trial design, such as when and how to use blocking designs, various methods of implementing randomization restrictions, and data analysis methodology (Cochran and Cox, 1957; Peterson, 1985; Quinn and Keough, 2002; Box et al., 2005; Hinkelmann and Kempthorne, 2008). Second, equipment dimensions, convenience, and personal preferences drive many decisions, particularly size and shape of experimental units and blocks. Many researchers are locked into particular dimensions that are partly determined by the size or capacity of planting or harvesting equipment, the size and shape of fields available for research, and colleagues' perceptions, or perhaps even peer pressure.

Perceptions play an important role in experimental design. On one hand, many researchers are under the perception that there is no efficient scientific method of predicting appropriate size of experimental units and blocks for field-based research. Instead, their decisions are based largely on their educational experiences, influences from colleagues conducting similar types of research, compilations from published studies of similar research, or guesses. On the other hand, many researchers are often reluctant to make significant changes to experimental design methodology, largely due to perceptions that they will be subject to criticism from colleagues and referees, reluctance to leave the current "comfort zone", or because they do not have the tools to predict the outcomes associated with changes in protocol. It is important to recognize that decisions we make with respect to experimental designs, for each and every field study that we establish, are not always active decisions, but may often be passive decisions. In other words, "If you choose not to decide, you still have made a choice" (Peart, 1980).

The objective of this paper is to summarize 28 years of field-based research at a single location describing changes to experimental design methodology, illustrating both predicted and empirical results of those changes, and demonstrating how some relatively simple computations and analytical methods can be used to predict the effects of change for any researcher at any location where a reasonable amount of historical data exists. These computations and predictions can be made from data collected on any routine field experiment, but the random nature of these variables lends significant doubt about the validity of extrapolating results from one field experiment to many future experiments. Rather, this paper uses data from 114 field experiments conducted at one location to establish trends from which predictions can be made for future experiments at this location, illustrating how this could be done for any quantitative variable measured routinely over many experiments conducted at any site. This paper illustrates a model that could be employed by any long-term field-based research program that aims to improve precision and efficiency of phenotypic evaluation in field-based experiments.

MATERIAL AND METHODS

EXPERIMENTAL MATERIALS AND DESIGNS

The study was conducted at the University of Wisconsin Arlington Agricultural Research Station (43.33° N, 89.38° W). Experiments were planted between 1981 and 2007 on a Plano silt loam soil (fine-silty, mixed, superactive, mesic Typic Argiudoll). The maximum slope was 1 to 2%. All trials followed 1 year of soybean [*Glycine max* (L.) Merr.] in a SGGGG or CSGGGG crop rotation, where S = soybean, C = corn (maize), and G = grass.

The study is based on an analysis of total forage yield for 114 genetic experiments of 11 forage grass species (Table 1): Kentucky bluegrass (*Poa pratensis* L.), meadow bromegrass (*Bromus riparius* Rehm.), meadow fescue [*Schedonorus pratensis* (Huds.) P. Beauv.], orchardgrass (*Dactylis glomerata* L.), perennial ryegrass (*Lolium perenne* L.), quackgrass [*Elymus repens* (L.) Gould], reed canarygrass (*Phalaris arundinacea* L.), smooth bromegrass (*B. inermis* Leyss.), tall fescue [*S. phoenix* (Scop.) Holub], timothy (*Phleum pratense* L.), and tall oatgrass [*Arrhenatherum elatius* (L.) P. Beauv. Ex J. Presl & C. Presl]. Treatments in each experiment consisted of cultivars, breeding lines, half-sib families, or accessions, always based on genetic variation as the underlying source of variation, hereafter termed "genetic lines". Five experimental designs were employed in these field experiments: randomized complete block (RCB); blocks in replicates, B/R (Casler, 1998); double simple lattice, DSL (Casler, 1999); augmented Latin Square, AUG (Casler et al., 2001) and split-plot randomization restriction within randomized complete blocks (SP).

Plots were planted with a drill planter in five rows spaced 15 cm apart using recommended seeding rates. All seed was planted on a pure-live-seed basis after adjusting for seed size and germination computed from in-house tests of all seed lots using standard procedures (AOSA, 2010). Plots were planted in late April and establishment-year management consisted of two or three clippings to control annual weeds without fertilization or data collection. Plot lengths were 1.5, 3.0, or 6.0 m and harvested plot areas were 1.4, 2.8, or 5.6 m^2 (Table 1).

Plots were fertilized with 160-275 kg N ha^{-1} year^{-1}, generally split equally among all harvests. The entire plot in 93 machine-harvested experiments was harvested with a flail-type harvester for the specified number of harvests in Table 1. The only exception was the quackgrass plots, which were planted in 10 drill rows (1.5 m wide) and harvests were made from a 0.9-m strip in the center of each plot. Dry matter determinations were made on random 300- to 500-g forage samples and were used to adjust plot yields to a dry matter basis. Dry matter yields for each plot were summed over all harvests within each year.

The remaining 21 experiments were planted on a different part of the research station, approximately 1 km apart from the machine-harvested experiments. These experiments were grazed with cattle, generally five times per year. Grazing was generally initiated when the canopy was between 20 and 30 cm in height. Forage yield (net forage available) was estimated immediately prior to each grazing event, using a rising plate meter calibrated at the end of the experiment to a set of samples, usually from n=100 to 200, representing each harvest over the duration of the experiment. For extremely small experiments, calibrations were pooled across species when found to be homogeneous (Casler et al., 1998). Each experiment was mowed to a residual height of 7 cm immediately after each grazing event so that “forage yield” would be measured on an equivalent basis for both grazed and machine-harvested experiments.

Statistical Analyses

Each experiment was analyzed by linear mixed models analysis, assuming blocks to be a random effect and both genetic lines and years to be fixed effects. For incomplete block designs (Table 1), both complete and incomplete blocks were assumed to have random effects. Residuals were computed and evaluated for normality using hypothesis tests in SAS PROC UNIVARIATE and visual inspection of normal probability plots. Residuals were generally normally distributed with very few exceptions. Residuals were plotted against predicted values to evaluate homogeneity of variance. Eighteen experiments had heterogeneous variances across years and this effect was modeled as distinct variance groups using the "repeated" statement in SAS PROC MIXED (Littel et al., 1996).

Table 1. Number of genetic lines (g), replicates (r), rows (R), columns (C), years (y), total number of harvests (h), and plot size (eu) for 114 perennial forage grass field experiments conducted between 1981 and 2009 at Arlington, WI.

Species[a]	Year[b]	Experiment[c]	Design[d]	g	r	R	C	y	h	eu
										m^2
KB	1995	GKB95B	RCB	2	4	2	4	3	15	5.6
KB	1995	KB95B	RCB	2	4	2	4	3	8	2.8
MB	1994	GMB94	RCB	4	4	4	4	3	15	5.6
MB	1994	MB94	RCB	2	4	2	4	3	8	2.8
MF	1996	GIPP	B/R	170	4	20	34	2	10	1.4
OG	1981	OG81	RCB	33	5	5	33	4	9	5.6
OG	1982	OG82	RCB	34	5	5	34	3	9	5.6
OG	1983	OG83	RCB	8	4	4	8	3	9	5.6
OG	1984	OG84	DSL	49	4	7	28	3	9	2.8
OG	1985	OG85	DSL	30	4	12	10	3	7	2.8
OG	1986	OG86	DSL	30	4	12	10	3	6	2.8
OG	1986	OPIE	B/R	448	2	56	16	2	5	1.4
OG	1987	OG87	RCB'	35	4	10	14	3	6	2.8
OG	1988	OG88	DSL	36	4	12	12	3	7	2.8
OG	1989	NE144F	SP	30	3	18	5	3	9	2.8
OG	1989	OG89	RCB'	56	4	8	28	3	9	2.8
OG	1991	OG91	RCB'	16	4	8	8	3	9	2.8
OG	1993	FIYS-OG	B/R	350	2	35	20	2	4	1.4
OG	1994	GOBE-E	AUG	25	3	15	5	3	15	5.6
OG	1994	GOBE-L	AUG	25	3	15	5	3	15	5.6
OG	1994	GOBE-M	AUG	36	3	18	6	3	15	5.6
OG	1994	HOBE-E	AUG	25	3	15	5	3	9	2.8
OG	1994	HOBE-L	AUG	25	3	15	5	3	9	2.8
OG	1994	HOBE-M	AUG	36	3	18	6	3	9	2.8
OG	1995	GOB95B	RCB'	10	4	5	8	3	15	5.6
OG	1995	GOG95A	RCB'	10	4	5	8	3	15	5.6
OG	1995	OG95A	RCB'	10	4	5	8	3	8	2.8
OG	1995	OG95B	RCB'	10	4	5	8	3	8	2.8
OG	1997	OG144	SP	30	4	24	5	3	9	2.8
OG	1998	FIYS2-OG	SP	16	16	32	8	2	6	1.4
OG	2002	PASSO	RCB	3	8	3	8	3	9	2.8
OG	2007	OPUS3F	RCB	7	4	7	4	2	10	2.8
OG	2007	OPUS3I	RCB	7	4	7	4	2	10	2.8
PR	1982	PR82	RCB	4	4	4	4	3	9	5.6
PR	1983	PR83	RCB	6	4	4	6	3	9	5.6
PR	1984	PR84	DSL	30	4	6	20	3	6	2.8
PR	1985	PR85	DSL	36	4	12	12	3	3	2.8
PR	1991	PR91	RCB'	32	4	16	8	3	9	2.8
PR	1994	GPR94	RCB	7	4	7	4	3	15	5.6
PR	1994	PR94	RCB	5	4	5	4	3	8	2.8
PR	1997	WFL97	RCB'	3	8	3	8	3	9	2.8
QG	1983	QUIER	B/R	350	2	70	10	2	4	1.4
QG	1988	SEER	RCB'	10	4	10	4	2	6	2.8
QG	1991	QG91	RCB	3	4	3	4	3	9	2.8
QG	1992	SEERR	RCB'	16	4	8	8	2	6	1.4
QG	1993	FIYS-HW	B/R	420	2	42	20	2	4	1.4
QG	1994	GQG94	RCB	2	4	2	4	3	15	5.6
QG	1994	QG94	RCB	2	4	2	4	3	8	2.8
QG	1998	FIYS2-HW	SP	16	16	32	8	2	6	1.4
RC	1984	RC84	DSL	16	4	8	8	3	9	2.8
RC	1989	RC89	RCB	3	4	3	4	3	9	2.8
RC	1994	GRC94	RCB	6	4	6	4	3	15	5.6
RC	1994	RC94	RCB	4	4	4	4	3	8	2.8
RC	1995	GRC95A	RCB	4	4	4	4	3	15	5.6
RC	1995	GRC95B	RCB	4	4	4	4	3	15	5.6
RC	1995	RC95A	RCB	4	4	4	4	3	8	2.8
RC	1995	RC95B	RCB	4	4	4	4	3	8	2.8
RC	2001	STORC	RCB'	15	8	20	6	2	7	1.4
RC	2002	PASSR	RCB	3	8	3	8	3	9	2.8
RC	2005	NARC	RCB'	88	3	33	8	2	6	1.4

SB	1982	SB82	RCB	13	4	13	4	3	6	5.6
SB	1985	LAND1	RCB'	32	3	12	8	2	6	1.4
SB	1989	LAND2	RCB'	40	3	15	8	2	6	1.4
SB	1991	SB144	RCB	30	4	24	5	3	7	2.8
SB	1991	SB91	RCB	3	4	3	4	3	9	2.8
SB	1992	NEGS	SP	18	3	9	6	3	8	2.8
SB	1993	FIYS-SB	B/R	350	2	35	20	2	4	1.4
SB	1994	GSB94	RCB'	20	4	10	8	3	15	5.6
SB	1994	SB94	RCB'	14	4	7	8	3	8	2.8
SB	1997	SB97N	SP	24	4	12	8	3	9	2.8
SB	1997	SB97P	SP	24	4	12	8	3	9	2.8
SB	1998	FIYS2-SB	SP	16	16	32	8	2	6	1.4
SB	1999	DIAT	SP	54	4	24	9	2	4	1.4
SB	2000	DINS	SP	28	4	14	8	3	6	2.8
SB	2001	BEGM2	SP	70	4	40	7	2	4	1.4
SB	2001	BEGM3	SP	70	4	40	7	2	5	1.4
SB	2001	BEGM4	SP	70	4	40	7	2	6	1.4
SB	2002	DINS2	SP	28	4	14	8	3	6	2.8
SB	2002	RHIFS	SP	48	4	24	8	3	8	2.8
SB	2003	DIAT2	SP	54	4	24	9	2	4	1.4
TF	1983	TF83	RCB	6	4	4	8	3	9	5.6
TF	1985	TF85	DSL	16	4	8	8	3	7	2.8
TF	1986	TF86	DSL	16	4	8	8	3	6	2.8
TF	1987	TF87	RCB'	14	4	14	4	3	6	2.8
TF	1989	TF89A	RCB'	12	4	8	6	3	8	2.8
TF	1989	TF89B	RCB'	12	4	6	8	3	9	2.8
TF	1991	TF91	RCB'	8	4	4	8	3	9	2.8
TF	1994	GTF94	RCB	6	4	6	4	3	15	5.6
TF	1994	TF94	RCB	4	4	4	4	3	8	2.8
TF	1995	GTF95A	RCB	5	4	5	4	3	15	5.6
TF	1995	GTF95B	RCB	5	4	5	4	3	15	5.6
TF	1995	TF95A	RCB	5	4	5	4	3	8	2.8
TF	1995	TF95B	RCB	5	4	5	4	3	8	2.8
TF	1996	FENDO	B/R	440	2	20	44	2	6	1.4
TM	1982	TM82	RCB	12	4	4	12	3	6	5.6
TM	1983	TM83	RCB	4	4	4	4	3	6	5.6
TM	1985	TM85	DSL	16	4	8	8	3	6	2.8
TM	1986	TM86	DSL	20	4	10	8	3	5	2.8
TM	1987	TM87	RCB'	28	4	14	8	3	6	2.8
TM	1988	TM88	RCB'	18	2	6	6	3	6	2.8
TM	1989	TM89	RCB'	12	4	6	8	3	9	2.8
TM	1991	TM91	RCB'	8	4	4	8	3	9	2.8
TM	1994	GTM94	RCB'	10	4	5	8	3	15	5.6
TM	1994	TM94	RCB	7	4	7	4	3	8	2.8
TM	1995	GTM95A	RCB	3	4	3	4	3	15	5.6
TM	1995	GTM95B	RCB	3	4	3	4	3	15	5.6
TM	1995	TM95A	RCB	3	4	3	4	3	8	2.8
TM	1995	TM95B	RCB	3	4	3	4	3	8	2.8
TM	1999	TIMPE1F	B/R	340	4	34	40	2	8	1.4
TM	1999	TIMPE1I	B/R	340	4	34	40	2	4	1.4
TM	2007	TIMPE2F	RCB	12	6	18	4	2	8	2.8
TM	2007	TIMBE2I	RCB	12	6	18	4	2	8	2.8
TO	1994	GTO94	RCB	3	4	3	4	3	15	5.6
TO	1994	TO94	RCB	3	4	3	4	3	8	2.8

[a] KB = Kentucky bluegrass, MB = meadow bromegrass, MF = meadow fescue, OG = orchardgrass, PR = perennial ryegrass, QG = quackgrass, RC = reed canarygrass, SB = smooth bromegrass, TF = tall fescue, TM = timothy, TO = tall oatgrass.

[b] Establishment year (no data collected during establishment year).

[c] Experiments with a name beginning in 'G' were grazed. All others were machine harvested.

[d] AUG = augmented Latin Square, B/R = blocks-in-reps design, DSL = double simple lattice, RCB = randomized complete block in which blocks and rows are identical (Design A of Casler, 1999), RCB′ = randomized complete block in which there are undesigned row blocks within complete blocks (Design B of Casler, 1999), SP = split-plot randomization restriction within randomized complete blocks.

Year effects were modeled as a repeated measures factor, using either compound symmetry or heterogeneous compound symmetry covariance structures (Littel et al., 1996). Separate random error terms were modeled for years, genetic lines, and year x genetic line as suggested by Steel et al. (1997). Year x genetic line interactions were not the focus of this experiment and were completely ignored in the analyses described herein. Regardless, year x genetic line interaction was significant in only about half of the experiments and it generally made up only about 10-30% of the variance of a "genetic line" mean in those experiments.

In addition to the linear mixed models analysis, each experiment was analyzed by nearest neighbor analysis using two covariate terms (mean of direct north-south neighbors and mean of direct east-west neighbors). Nearest neighbor analyses were conducted as described by Casler (1999) and Smith and Casler (2004). The two covariates were fitted as random effects, including interaction terms with "years" in those rare cases where Akaike's Information Criterion suggested a better fit to the model, accounting for differential spatial variation across years (Littel et al., 1996).

The coefficient of variation, coefficient of heterogeneity, relative efficiency of blocking, and relative efficiency of nearest neighbor analysis was computed for every experiment. The CV was computed as CV = $100(r)(V_m)/M$, where r = the number of replicates and M = the estimated grand mean of the experiment. Relative efficiencies of Design 2 (Analysis 2) relative to Design 1 (Analysis 1) were computed as RE = $100(V_{m1}/ V_{m2})$ and adjusted for degrees of freedom when $df_{e1} < 20$ (Steel et al., 1997), where V_m = the average variance of a genetic line least-squares mean, accounting for variation associated with all random effects, including estimation errors of nearest neighbor covariates. For relative efficiency computations, Design 2 was the design employed, while Design 1 was the next simpler design for comparative purposes, e.g. randomized complete block vs. completely randomized design, lattice design vs. randomized complete block design.

For experiments with 5.6- or 2.8-m^2 plots, formulas presented by Lin and Binns (1984) were used to predict the resulting CV if plot size was reduced by 50%. This computation was achieved in four steps. First, estimate the block variance component as

$$s^2_B = (MS_B – MS_e)/t$$

where t = number of treatments and the two mean squares are for blocks and the error term relevant to treatments (genetic lines) and MS = mean square. These values were estimated as random effects using restricted maximum likelihood estimation within linear mixed models analysis (Littel et al., 1996). This computation was conducted on the complete-block source of variation for every experiment (all designs, including all incomplete block designs, had a random effect associated with complete blocks). Second, compute the intrablock correlation as

$$r_I = s^2_B/(s^2_B + s^2_e)$$

where s^2_e = the residual variance from linear mixed models analysis (e.g. block × year × genetic line interaction for the RCB design). Third, compute the coefficient of intrablock heterogeneity (Smith, 1938) as

$$b = 1 – \log[t – (t -1)(1 - r_I)]/\log(t).$$

Fourth, the expected CV for a 50% reduction in plot size was computed for all experiments with 5.6- or 2.8-m^2 plots as

$$CV_{exp} = CV_{obs}(10^{-b[\log(0.5)/2]})$$

where CV_{obs} is the observed CV from each experiment (Lin and Binns, 1984).

One-way analysis of variance was used to analyze data for CV_{obs}, CV_{exp}, b, and RE. Fixed effects were fitted to these variables to account for species, establishment year, harvest method (grazing vs. machine), design type, number of genetic lines, the log-linear effect of plot size and all first-order interactions between these effects.

Lastly, a series of half-sib progeny tests (FIYS-OG, FIYS-HW, and FIYS-SB experiments in Table 1; Casler, 1998) were used to generate a power function for detection of significant differences among genetic lines using P(Type I error) = 0.05. The purpose of this computation was to predict the number of replicates required to detect desired differences in future experiments designed to evaluate progeny populations selected from these three experiments. The minimum number of replicates required was computed as

$$r = [(t_\alpha + t_\beta)CV/d]^{0.5}$$

where t_α and t_β are two-tailed Student's t-values for Type I (α) and Type II (β) error rates, and d = the desired detection level between treatment means as a percentage of the experiment mean (Steel et al., 1997). Based on the power function, r = 16 replicates were chosen for the next series of experiments (FIYS2-OG, FIYS2-HW, and FIYS2-SB experiments in Table 1; Casler, 2008) and a comparative analysis was conducted to assess the effectiveness of the power function to predict desired detection levels.

RESULTS AND DISCUSSION

Coefficients of variation ranged from 3.1 to 43.3% (Fig. 1). Coefficients of heterogeneity ranged across nearly the entire range of possible values, from 0.045 to 1.000, but were decidedly skewed toward higher values (Fig. 2). Relative efficiency of blocking ranged from 89 to 321% for incomplete block designs compared to the randomized complete block design (Fig. 3). Relative efficiency of nearest neighbor analysis ranged from 89 to 349% compared to the randomized complete block design without spatial analysis (Fig. 3).

Grass species and establishment year had a significant effect on coefficient of heterogeneity, coefficient of variation, and relative efficiency of NNA (Table 2). Experimental design had a significant effect only on CV, while experiment size (number of genetic lines) had no impact on these three statistics. There was considerable variation among and within species for all three of these statistics (Table 3). Perennial ryegrass had the highest average b-values and CV values, probably due to its role as the least winterhardy of these species at this location. High values of b and CV are indicative of ineffective blocking and high levels of unexplained variation, both of which can be symptomatic of variability in forage yield induced by large and unpredictable stand losses. While there was significant variation among experimental design types for CV, complex randomizations and incomplete block designs with small block sizes, e.g. lattice designs, were not a guarantee against high CV values (Table 4). Augmented Latin Squares were the most efficient design, following by the various split-plot randomizations of the RCB design. Coefficients of heterogeneity, observed CVs, and predicted CVs were significantly greater for machine-harvested plots compared to grazed plots, all with a common plot size of 5.6 m^2 (Tables 2 and 5). Reducing plot size was expected to have a greater impact on CV for machine-harvested plots (30.6% increase) compared to grazed plots (17.4% increase). Establishment year had the largest effect of all these fixed effects, with ranges of b = 0.31 ± 0.25 to 1.00 ± 0.00, CV = 5.2 ± 0.9 to 21.4 ± 13.2, and RE(NNA) = 100 ± 3 to 270 ± 27 (data not shown). There were no significant effects of any first-order interactions among these fixed effects.

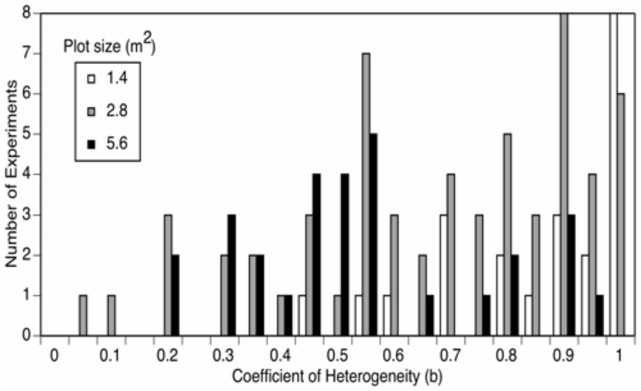

Figure 1. Frequency distribution of coefficients of variation for mean forage yield of 114 perennial forage grass experiments conducted at Arlington, WI between 1981 and 2009.

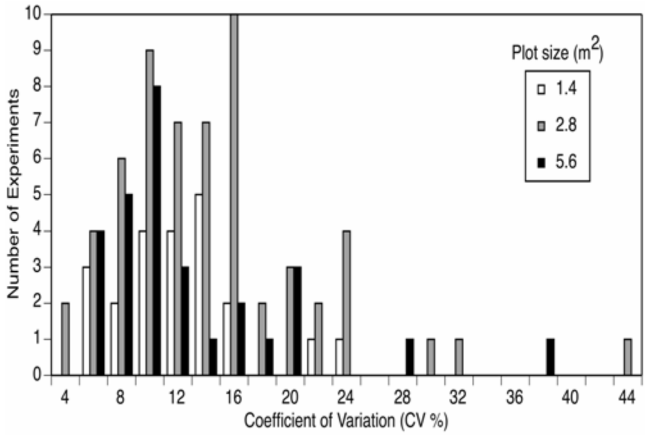

Figure 2. Frequency distribution of coefficients of heterogeneity for mean forage yield of 114 perennial forage grass experiments conducted at Arlington, WI between 1981 and 2009.

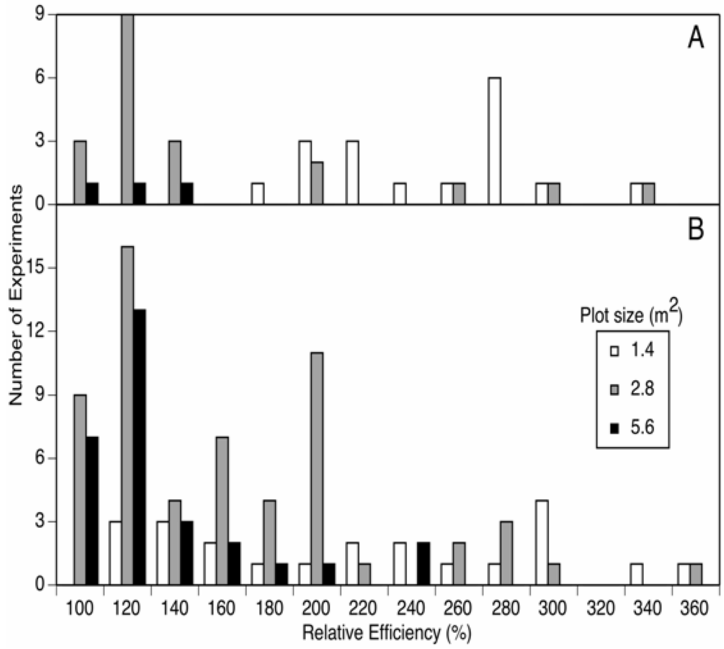

Figure 3. Frequency distribution of relative efficiencies for A) incomplete blocking or B) nearest neighbor analysis, based on analysis of variance of forage yield for 114 perennial forage grass experiments conducted at Arlington, WI between 1981 and 2009.

Table 2. *P*-values of fixed effects associated with estimated coefficients of heterogeneity (b), coefficients of variation (CV), and relative efficiency of nearest neighbor analysis [RE(NNA)] for 114 perennial forage grass experiments planted between 1981 and 2007.

Effect	df	b	CV	RE(NNA)
Species	10	0.0148	0.0002	0.0103
Establishment year	23	0.0045	0.0306	0.0123
Grazing vs. Machine harvesting	1	<0.0001	<0.0001	0.4747
Experimental design	5	0.4111	0.0276	0.0918
Number of genetic lines	1	0.6995	0.9793	0.5838
Plot size (log linear)	1	0.1624	0.0101	<0.0001

Table 3. Mean plus or minus standard deviation for coefficients of heterogeneity (b), coefficients of variation (CV), and relative efficiency of nearest neighbor analysis [RE(NNA)] estimated on seven of 11 species with at least eight experiments (n) as shown in Table 1.

Species	n	b	CV	RE(NNA)
Orchardgrass	28	0.63 ± 0.23	12.2 ± 5.6	155 ± 79
Perennial ryegrass	8	0.84 ± 0.16	23.0 ± 13.1	144 ± 1
Quackgrass	8	0.74 ± 0.21	10.0 ± 5.6	152 ± 60
Reed canarygrass	11	0.64 ± 0.26	10.8 ± 3.5	173 ± 38
Smooth bromegrass	20	0.66 ± 0.21	10.4 ± 3.3	197 ± 77
Tall fescue	14	0.61 ± 0.26	15.4 ± 6.5	141 ± 7
Timothy	16	0.59 ± 0.26	12.9 ± 4.7	115 ± 99

Table 4. Mean plus or minus standard deviation for coefficients of variation (CV) estimated from six types of experimental designs employed in 114 perennial forage grass field experiments, as shown in Table 1 (n = number of experiments).

Experimental design	n	CV
Augmented Latin Square design	6	4.7 ± 2.2
Blocks in Reps design	8	13.0 ± 5.9
Lattice designs	11	21.7 ± 9.7
Randomized complete block (RCB)[a]	45	11.3 ± 5.9
Randomized complete block (RCB')[a]	28	13.4 ± 5.6
Split plots	14	9.7 ± 2.7

[a] RCB = randomized complete block in which blocks and rows are identical (Design A of Casler, 1999), RCB′ = randomized complete block in which there are undesigned row blocks within complete blocks (Design B of Casler, 1999).

Table 5. Empirical mean coefficients of heterogeneity and coefficients of variation for grazed and machine-harvested plots with a common plot size of 5.6 m^2, including predicted coefficients of variation for decreasing plot size from 5.6 to 2.8 m^2. The underlying measurement variable is forage yield of 29 perennial forage grass experiments conducted at Arlington, WI between 1981 and 2009.

Plot type	Number of experiments	Empirical mean coefficient of heterogeneity (*b*)	Empirical mean coefficient of variation (*CV*)	Mean predicted coefficient of variation (CV_{pred})
			%	%
Grazed	20	0.42	8.6	10.1
Machine-harvested	9	0.72	18.6	24.3
P-value		<0.0001	<0.0001	<0.0001

Table 6. Empirical mean coefficients of heterogeneity and coefficients of variation for three plot sizes and predicted coefficients of variation for decreasing plot size from 5.6 to 2.8 m^2 or from 2.8 to 1.4 m^2. The underlying measurement variable is forage yield of 94 perennial forage grass machine-harvested experiments conducted at Arlington, WI between 1981 and 2009.

Plot size [a]	Number of experiments	Empirical mean coefficient of heterogeneity (b)	Empirical mean coefficient of variation (CV)	Mean predicted coefficient of variation (CV_{pred})
			%	%
5.6 m^2	9	0.72	18.6	NA
2.8 m^2	63	0.64	13.3	24.3
1.4 m^2	22	0.83	11.5	16.9
Slope		NA	0.116	NA
P-value		0.1624	0.0101	NA
$R^2_{(3)}$		0.06	0.99	NA
$R^2_{(94)}$		0.01	0.07	NA

[a] Slope = linear regression coefficient of variable as a function of plot size. P-value determined from log-linear plot-size contrast in one-way ANOVA. $R^2_{(3)}$ = proportion of variation among plot-size means explained by regression. $R^2_{(94)}$ = proportion of variation among raw data explained by regression.

Table 7. Mean relative efficiencies of nearest neighbor analysis and incomplete blocking for 114 perennial forage grass experiments, including analysis of log-linear regressions on plot size. The underlying measurement variable is forage yield of 114 perennial forage grass experiments conducted at Arlington, WI between 1981 and 2009.

Plot size [a]	Number of experiments	Empirical mean relative efficiency of nearest neighbor analysis	Empirical mean relative efficiency of blocking [b]
		%	%
5.6 m^2	29	130	109
2.8 m^2	63	152	145
1.4 m^2	22	212	240
Slope		-0.112	-0.170
P-value		<0.0001	<0.0001
$R^2_{(3)}$		0.89	0.81
$R^2_{(114)}$		0.17	0.40

[a] Slope = linear regression coefficient of variable as a function of ln(plot size). P-value determined from log-linear plot-size contrast in one-way ANOVA. $R^2_{(3)}$ = proportion of variation among plot-size means explained by regression. $R^2_{(114)}$ = proportion of variation among raw data explained by regression.

[b] Only 40 trials qualified for estimation of the relative efficiency of incomplete blocking (n = 17, 20, and 3 with 5.6-, 2.8-, and 1.4- m^2 plot size).

Table 8. Least significant differences, expressed as a percentage of the mean, between genetic selections evaluated in 16-replicate field experiments at three locations in Wisconsin, USA (Experiments FIYS2-OG, FIYS2-SB, and FIYS2-HW in Table 1; data taken from Casler, 1999).

Location	Year	Number of observations	Orchardgrass	Smooth bromegrass	Hybrid wheatgrass
			%	%	%
Arlington	1999	16	5.7	6.3	5.9
Ashland	1999	16	5.6	4.7	3.4
Marshfield	1999	16	6.0	5.4	4.6
Arlington	2000	16	9.0	4.9	4.9
Ashland	2000	16	5.5	7.3	8.7
Marshfield	2000	16	3.8	4.8	4.9
Arlington	2001	16	5.3	4.9	5.7
Ashland	2001	16	6.8	8.6	10.6
Marshfield	2001	16	4.5	5.4	6.4
Arlington	3-yr mean	48	3.6	3.2	2.7
Ashland	3-yr mean	48	3.2	3.2	3.4
Marshfield	3-yr mean	48	3.1	3.6	4.2

The nine largest CV values and 26 of the 28 largest CV values originated from experiments with plot sizes of 5.6 or 2.8 m^2. The log-linear plot size effect was significant for CV (Table 2) and simple linear regression revealed a significant log-linear effect of plot size on empirical CV with a nearly perfect fit to the three plot-size means (Table 6). Due to the large amount of variation within plot sizes, the log-linear regression accounted for only 7% of the variation among raw CVs, but 99% of the variation among plot-size means. The positive slope was unexpected, because predicted CV values averaged a 30.6% increase for decreasing plot size from 5.6 to 2.8 m^2 and a 27.1% increase for decreasing plot size from 2.8 to 1.4 m^2, based on Smith's Law. Empirical CVs decreased by 28.5% and 13.5% for the two plot size reductions, respectively.

Higher values of b tend to result when blocking is relatively ineffective and are suggestive of situations in which CV is highly responsive to changes in plot size (Lin and Binns, 1984, 1986). There appeared to be a slight trend toward smaller plots having larger coefficients of heterogeneity (Fig. 2), but this relationship was not significant due to extreme variability within plot sizes (Tables 2 and 6). Nevertheless, of the 21 experiments with a plot size of 1.4 m^2, 20 had $b > 0.5$ for a frequency of 0.95; of the remaining 93 experiments with larger plot size, 61 had $b > 0.5$ for a frequency of 0.66. The contingency chi-square test ($\chi^2 = 7.32$; $p < 0.01$) indicated that these two frequencies were significantly different.

For the 40 incomplete block experiments, the relative efficiencies of incomplete blocking and nearest neighbor analysis were positively correlated with each other ($r = 0.65$; $p < 0.01$). Both relative efficiencies decreased significantly in a log-linear manner with increasing plot size (Table 7; Fig. 4). The log-linear model fit very well for plot-size means of both relative efficiencies. The response to increasing plot size was greater for incomplete blocking compared to nearest neighbor analysis.

Averaged across the three large experiments planted in 1993 (three species), the power function predicted that r = 16 replicates would detect differences of 3% of the mean with Power = 0.1, 4% of the mean with Power = 0.3, 5% of the mean with Power = 0.5, and 6% of the mean with Power = 0.7 (Fig. 5). Averaged across three years of data collection, increasing the number of effective “replicates” to 48, detection levels increased to 1% of the mean with Power = 0.65 and 2% of the mean with Power = 0.9. The three experiments planted in 1998, as Phase 2 of that project, with r = 16 replicates had detection levels ranging from 4.9 to 9.0%

of the experiment mean for individual years and 2.7 to 3.6% of the mean for means over years (Table 5). These values were indicative of Power = 0.50 to 0.85 for individual years and Power = 0.90 to 0.99 for means over years. Detection levels were similar for identical experiments planted at two additional locations (Ashland and Marshfield, WI). These values, deriving from the extreme high levels of replication, were critical in the analysis of data from these experiments, allowing detection of small but consistent differences between different breeding methodologies (Casler, 2008).

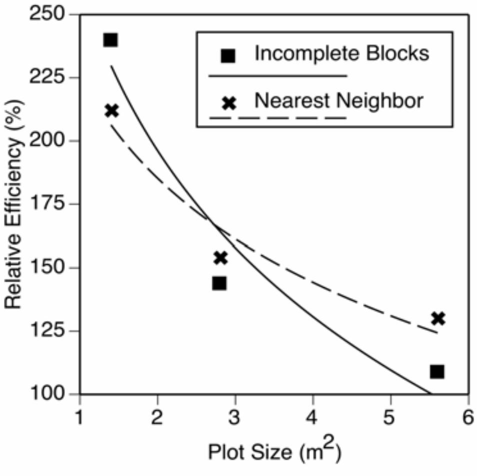

Figure 4. Log-linear regressions of mean relative efficiencies for nearest neighbor analysis and incomplete blocking as a function of plot size for perennial forage grass experiments conducted at Arlington, WI between 1981 and 2009. Statistics of these regression lines are shown in Table 4.

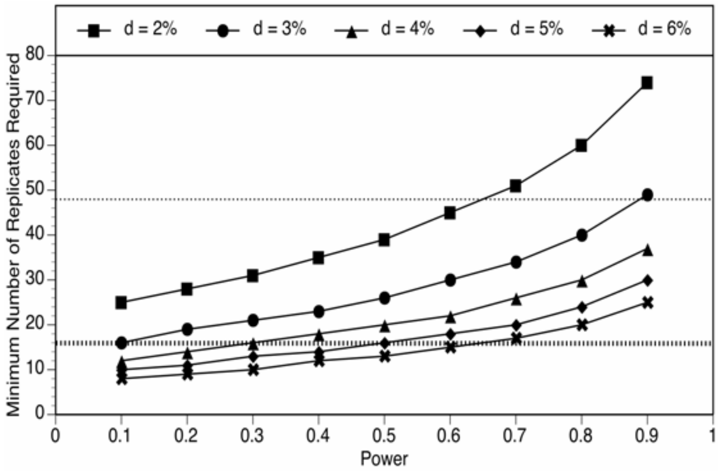

Figure 5. Power function illustrating the minimum number of replicates required to detect differences between genetic lines of d = 2 to 6% of the mean, with computations based on the coefficients of variation for experiments FIYS-OG, FIYS-HW, and FIYS-SB (Table 1). Computations were based on an average CV = 11.3% for the three field experiments and a Type I Error rate of $\alpha = 0.05$.

DISCUSSION

The variance among plots of a constant size x (V_x) was long ago shown to have a logarithmic relationship to plot size, expressed as

$$\log(V_x) = \log(V_1) - b[\log(x)],$$

where $\log(V_1)$ is the intercept of the log-linear regression (Smith, 1938). Because CV is simply a scaled version of V_x (strictly speaking, the trial mean is a constant with respect to plot size), it follows a similar relationship of an asymptotically decreasing function. When plot size is small, CV is highly responsive to changes in plot size - small increases in plot size should have a significant impact on reducing CV, but small decreases can be problematic. Conversely, when plot size is large, CV is relatively unresponsive to changes in plot size, suggesting that large reductions in plot size could be used to create more cost-efficient or space-efficient designs.

This relationship has always left researchers with a central core of questions: (i) what is the slope of this relationship for my field sites; (ii) can I afford to reduce plot size without sacrificing precision; and (iii) would I benefit significantly by increasing plot size? Prior to 1984, empirical plotting and regression analysis was the only mechanism to estimate the slope of this relationship (Koch and Rigney, 1951; Smith, 1938). This requires the laborious process of conducting a factorial or nested uniformity trial in which V_x is empirically estimated for a range of plot sizes (e.g. Casler and Tageldin, 1996). One cannot assume a particular value, because the slope of this relationship is highly variable, with values of b spanning nearly the entire range ($0 \leq b \leq 1$) in both the current study conducted at a single site and in Smith's study that spanned a wide array of sites and species.

Lin and Binns (1984) solved this problem by describing a simple set of computations that could be made from an ANOVA of any blocking design. These computations can be easily automated in a spreadsheet, requiring minimal routine input values: number of treatments and variance component estimate for blocks, which is now part of the routine output in linear mixed models and generalized linear mixed models analyses (Littel et al., 1996). The range in estimated coefficients of heterogeneity observed in this 28-yr case study (b = 0.04 to 1.00) clearly indicated high levels of instability for a group of similar species evaluated in trials conducted within a 6-ha area of one research station. Published results from uniformity trials have previously suggested that these estimates are highly stable across years within narrow time frames (e.g. Koch and Rigney, 1951; Casler and Tageldin, 1996). Of all the factors that varied among the 114 field studies in this case study, only one, grazed vs. machine harvested, accounted for any significant or consistent difference in estimated b values. Because there were no relationships of these values with species, establishment year, or field divisions, most of the inconsistency must be taken as inherent variability in these estimates. This instability suggests that numerous estimates of the coefficient of heterogeneity are required to obtain a reasonable assessment of the heterogeneity characteristics of a particular site, assessed by the distribution and range of possible values and the most likely values encountered (e.g. Fig. 2), as opposed to the simple mean across trials.

The failure of the Lin and Binns (1984) method to predict empirical changes in CV should not be viewed as an indication of inadequacy of this methodology. Rather, the decrease in CV associated with a *decrease* in plot size is a reflection of synergistic effects between changes to both plot size and experimental design that could not be predicted from these computations. Neither blocking nor spatial analyses were highly effective for the larger plot sizes utilized during this time period. However, numerous types of incomplete block designs, as well as nearest neighbor analyses, were highly effective for the smallest plot size, suggesting that soil heterogeneity is largely manifested on a very small scale within this experimental area. Incomplete block sizes generally ranged from 4 to 10 plots per block, for a total block size of 5.6 to 14.0 m^2 for 1.4-m^2 plots. Block sizes larger than these were marginally or sporadically effective, as shown by relatively low mean relative efficiencies (Table 4). Nearest neighbor analyses were sporadically effective for the largest plot size because of their inability to capture spatial variability on a fine scale. The larger plots were too internally heterogeneous to be effective in capturing spatial variability on this fine scale.

This result raises an interesting question: does Smith's Law of Heterogeneity apply to this site? Strictly speaking, yes. A classical uniformity trial conducted on one field within this 6-ha area behaved exactly as expected, with a large decrease in CV as a function of increasing plot size (Casler and Tageldin, 1996). In the broader sense, over time and space, the general principle of Smith's Law does not appear to be applicable to this particular site. The large amount of variability among estimates of heterogeneity coefficients and the lack of explanatory factors for that variability, suggest that the patterns of variability are random, unpredictable, unrepeatable, and manifested on a fairly small geographic scale. Incomplete block designs were considerably and consistently more effective than predicted by Casler and Tageldin (1996), suggesting that a single uniformity trial, as reliable as it may be for that particular point in space and time, cannot represent variability in neighboring fields, as similar as they may appear based on visual evaluations and soil-test data, nor long-term trends for that particular field.

Finally, the nature of spatial variability at this site suggests that routine use of the smallest practical and realistic plot size, sufficient to provide reliable estimates of sward-plot biomass yield, will most efficiently utilize scarce research resources. Ignoring cost for the moment, resource-allocation exercises using variance component analyses indicate that more replicates of smaller plots are always better than the converse when land area or seed quantities are fixed (Lin and Binns, 1984, 1986; Cherney et al., 1995; McCann et al., 2012). This is often the case in a breeding program where land area, labor, daylight hours, and seed quantities are always factors limiting the number of families and replicates that can be employed to conduct progeny tests (Casler and Brummer, 2008). For one series of experiments, power analysis was a highly reliable method of predicting the number of replicates required to detect a desired difference between treatment means. While some of the experiments utilized in this case study were strictly intended to "pick the winners" (to chose the best families for selection and breeding), most were intended to provide treatment means with LSD values to be used in decision analysis of significant and meaningful differences among treatments. Future field studies of this nature would benefit greatly from some preplanned thought about desired detection levels and a thorough power analysis to predict the number of replicates required to reach that detection level. These analyses may lead to widely varying decisions regarding the optimal number of replicates in field trials, a result that may be unsettling to some researchers who are accustomed to long-term routines.

The specific choice of experimental designs for this site does not appear to be critical for field experiments using plot sizes on the low end of the scale studied herein. Blocking experiments with quite small block sizes (4 to 12 plots) are a wise choice, providing two options for controlling spatial variation during the data analysis phase (allocating degrees of freedom and variation to blocks or conducting spatial analyses such as nearest neighbor analysis). Viable options include split-plots, lattices, row-column designs, blocks-in-reps (and reps-in-blocks), and numerous balanced and unbalanced incomplete block designs (Cochran and Cox, 1957; Peterson, 1985; Hinkelmann and Kempthorne, 2008). For small experiments, with fewer than 10 treatments, the penalty in lost degrees of freedom is potentially severe, suggesting that incomplete block designs should not be an automatic choice. Classically, completely randomized designs have not been frequently employed in field-based research. However, their simplicity and maximum error degrees of freedom, combined with modern and advanced methods of spatial analysis (Fischer and Getis, 2010) suggest that this design is a viable option for certain circumstances in which blocking designs are not desirable or practical.

That said, I offer one very important caveat for the routine use of blocking designs in this type of field-based research program. Well-placed and regularly spaced borders between blocks can be strategically used to control variability associated with unexpected events, such as equipment breakdowns, weather delays, or lack of labor availability. Any of these factors can cause delays in harvesting in the middle of medium to large field experiments. Because such delays result in discrete boundaries, some of which can be planned to match block boundaries, simple analyses that account for these effects as a fixed environmental effect, or as part of a random block effect, will be much more effective than trying to capture this variation in a continuous spatial variable across the entire experimental area. The advantage of preplanning block boundaries into the field experiment is that blocks can be organized to be orthogonal with treatments, whereas the introduction of unexpected delays or breaks at random places in the experiment would lead to potential lack of orthogonality. Numerous studies have demonstrated the empirical value of blocking designs, particularly those with smaller block sizes, such as 8-12 treatments (Baird and Mead, 1991; Lin et al., 1993; Kempton et al., 1994; Handa et al., 1995). In particular two-dimensional blocking can be extremely effective when little or nothing is known in advance about spatial variability in the field (Lin et al., 1993; Kempton et al., 1994). Both smaller block sizes and two-dimensional blocking lend themselves much more readily to retrospective adjustments of unexpected events, such as harvester breakdowns or weather delays.

Conclusions

Secondary statistics, computed from analyses of variance, can have value in predicting the effects of changes in experimental designs. However, the highly variable nature of these statistics across time and space indicate that long-term observations and trends are highly superior to short-term observations from a small number of field experiments. Over the 28-year history of these experiments, the decisions to reduce plot size from 5.6 to 2.8 m^2 and from 2.8 to 1.4 m^2 were made strictly on a practical basis, due to limitations in land area available, seed quantities, or equipment size and functionality. The reductions in CV associated with reduced plot size, and concomitant increases in power, were completely serendipitous, owing to the nature of spatial variation that could only be adequately described in this retrospective analysis for this particular site. Conversely, the reductions in CV associated with increasing the number of replicates were largely as predicted by power analyses, indicating the value of this underutilized method for assisting in the design of experiments. While the conclusions drawn from this retrospective analysis apply strictly to this site, there are likely numerous other sites that share similar characteristics. A similar retrospective analysis would reveal these characteristics, but also be of value for sites with different spatial characteristics, potentially revealing unknown patterns and the scale at which spatial variation interferes with the ability to detect differences among treatment means.

References

AOSA. (2010). *Association of Seed Analysts, Rules for Testing Seeds.* http://www.aosaseed.com/publications.htm

Baird, D., Mead, R. (1991). The empirical efficiency and validity of two neighbor models. *Biometrics* 47, 1473–1487.

Box, G.E.P., Hunter, J.S., Hunter, W.G.. (2005). *Statistics for experimenters: Design, innovation, and discovery.* 2nd ed. Wiley-Interscience, NY

Casler, M.D. (1998). Genetic variation within eight populations of perennial forage grasses. *Plant Breeding* 117, 243–249.

Casler, M.D. (1999). Spatial variation affects precision of perennial cool-season forage grass trials. *Agronomy Journal* 91, 75–81.

Casler, M.D. (2008). Among-and-within-family selection in eight forage grass populations. *Crop Science* 48, 434–442.

Casler, M.D. and Brummer, E.C. (2008). Theoretical expected genetic gains for among-and-within-family selection methods in perennial forage crops. *Crop Science* 48, 890–902.

Casler, M.D., Fales, S.L., Undersander, D.J., and McElroy, A.R. (2001). Genetic progress from 40 years of orchardgrass breeding in North America measured under management intensive rotational grazing. *Canadian Journal of Plant Science* 81, 713–721.

Casler, M.D., and Tageldin, M.H. (1996). Experimental design factors affecting error variation in orchardgrass. *Agronomy Journal* 88, 745–749.

Casler, M.D., Undersander, D.J., Fredericks, C., Combs, D.K., and Reed, J.D.. (1998). An on-farm test of perennial forage grass varieties under management intensive grazing. *Journal of Production Agriculture* 11, 92–99.

Cherney, J.H., Casler, M.D., Cherney, D.J.R. (1996). Sampling forage corn for quality. *Canadian Journal of Plant Science* 76, 93–99.

Cochran, W.G., Cox, G.M. (1957). *Experimental Designs.* John Wiley & Sons, Inc. NY.

Fischer, M.M., Getis, A., (eds.) (2010). Handbook of applied spatial analysis: Software tools, methods, and applications. Springer, NY.

Handa, D.P., Sreenath, P.R. , Rajpali, S.K. (1995). Uniformity trial with lucerne grown for forage. *Grass and Forage Science* 50, 209–216.

Hinkelmann, K., Kempthorne, O. (2008). *Design and analysis of experiments. I. Introduction to experimental design*. 2nd ed. Wiley-Interscience, NY.

Kempton, R.A., Seraphin, J.C., Sword, A.M. (1994). Statistical analysis of two-dimensional variation in variety yield trials. *Journal of Agricultural Science* 122, 335–342.

Koch, E.J., Rigney, J.A. (1951). A method of estimating optimum plot size from experimental data. *Agronomy Journal* 43, 17–21.

Lin, C.S., Binns, M.R. (1984). Working rules for determining the plot size and numbers of plots per block in field experiments. *Journal of Agricultural Science* 103, 11–15.

Lin, C.S., Binns, M.R. (1986). Relative efficiency of two randomized block designs having different plot sizes and numbers of replications and plots per block. *Agronomy Journal* 78, 531–534.

Lin, C.S., Binns, M.R., Voldeng, H.D., Guillemette, R. (1993). Performance of randomized block designs in field experiments. *Agronomy Journal* 85, 168–171.

Littel, R.C. Milliken, G.A., Stroup, W.W., Wolfinger, R.D. (1996). *SAS System For Mixed Models.* SAS Institute, Inc. Cary, NC.

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

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

Peterson, R.G. (1985). *Design and analysis of experiments.* Marcel Dekker, Inc., NY.

Quinn, G.P., Keough, M.J. (2002). *Experimental design and data analysis for biologists.* Cambridge University Press, Cambridge, UK.

Smith, H.F. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *Journal of Agricultural Science* 28, 1–23.

Smith, K.F., Casler, M.D. (2004). The use of spatially adjusted herbage yields during the analysis of perennial forage grass trials across locations. *Crop Science* 44, 56–62.

Steel, R.G.D., Torrie, J.H., Dickey, D.A. (1997). *Principles and procedures in statistics.* 3rd ed. McGraw-Hill.