1998

Statistical Design and Analysis of Producer/ Consumer Evaluations to Assess Plant Quality

Walter W. Stroup
University of Nebraska - Lincoln, wstroup1@unl.edu

Stacy A. Adams
University of Nebraska - Lincoln

Ellen Paparozzi
University of Nebraska - Lincoln

Statistical Design and Analysis of Producer/Consumer Evaluations to Assess Plant Quality

W.W. Stroup[1]
Department of Biometry, University of Nebraska, 103 Miller Hall, Lincoln, NE 68583-0712

Stacy A. Adams[2] **and Ellen T. Paparozzi**
Department of Horticulture, University of Nebraska, 377 Plant Sciences, Lincoln, NE 68583-0724

Researchers in ornamental horticulture often want to assess the effects of experimental treatments on plant quality. Producers often use the results of such experiments to establish the minimum level of a treatment, for instance, the amount of a growth regulator or a nutrient, such as nitrogen, needed to achieve desired plant quality. For edible plants, quality can be quantified objectively by using numeric response variables such as yield or nutritional content. However, for ornamental plants, quality depends on aesthetic appeal and consumer acceptance, traits which are subjective and qualitative.

Statistical methods for the design and analysis of experiments involving numeric or quantitative responses are generally considered "standard" statistical methods. Often, however, the relationship between these variables and quality factors such as aesthetic appeal and consumer acceptance is not clear. For subjective, qualitative response variables, standard statistical methods may not be used without modification. The purpose of this paper is to present statistical methods useful for designing and analyzing experiments to assess plant quality. We focus particularly on the use of unreplicated designs and their analysis using half-normal plots.

SPECIAL ASPECTS OF QUALITY ASSESSMENT

Specific aspects of experimental design and analysis in quality assessment experiments set them apart from research using quantitative variables. These can be divided into 1) requirements and constraints that must be reflected in the design, and 2) special considerations that must be reflected in the analysis.

Considerations and constraints on the design. A typical objective of a quality assessment experiment is to determine minimum required level of a treatment needed to achieve the desired plant response and the desired level of consumer acceptance. Producers face pressure to minimize inputs for both economic and environmental reasons, while at the same time preserving quality. A "good" design is one that allows this objective to be addressed as accurately and efficiently as possible.

For efficiency, the design should use the fewest levels of a treatment possible without compromising accuracy. Because the effect of the level of treatment applied on quality may not be well known, a wide range of levels should be tested, including the minimum and maximum treatment level considered plausible. In addition, the increments between levels should be small, to insure that the optimum level is found. If the levels are too far apart and the optimum is midway between, the experiment will result in a substantial overestimate of the optimum. While there are no hard and fast rules for the required number of levels and their spacing, our experience indicates that six to eight levels per treatment are sufficient to permit accurate estimation of the optimum. Criteria for choosing treatment levels are discussed in texts such as Box et al. (1978) and Mead (1988).

Quality assessment requires rating plants on a subjective scale. To be meaningful, a scale must satisfy two basic requirements. First, the number of categories in the scale must be sufficient to distinguish among treatments, but not so numerous as to be confusing. Second, the categories used must have tangible meaning to individuals doing the rating.

Selecting the number of categories requires judgment; there are no exact rules. Agresti (1990) and Freeman (1987), for example, recommend minimizing the number of categories, by using two, or at most three, when possible. They warn that as the number of categories increases, the models required for analysis become increasingly difficult to interpret. They also caution that peoples' ability to make meaningful classifications declines as the number of categories increases. Cloninger et al. (1976), in their study of sensory evaluation methods, recommend using rating scales with five categories. Their work suggests that more categories are confusing. Other studies, for example, Cox (1980), suggest using more categories, which, in theory, allow greater sensitivity to treatment differences. All studies stipulate that the categories must be clear to the individuals doing the rating.

In practice, the rating scale reflects the objective of the study. In many cases, binary-scales, e.g., "like it and would buy it" or "don't like it," are sufficient. Three-point scales, e.g., "unacceptable," "acceptable for a discount store," or "florist quality," are more suitable for other applications. The categories should not be stated in jargon, but should be meaningful to the rater. For example, "acceptable" vs. "unacceptable" may be meaningful to producers, but the fundamental question is whether consumers like the plant enough to buy it. "Like it and would buy it" vs. "don't like it" focuses their attention without any ambiguity.

The need for subjective ratings conflicts with the need for many treatment levels. This is because people can rate only so many plants without tiring. Stone and Siedel (1993) discuss various sources of rating error, including the effect of tiring, often called "rater fatigue." The design must not require people to rate so many plants that they become fatigued. Specific ways of doing this are discussed in the next section.

To summarize, the three main design considerations are:

1. A range of treatment levels that includes the minimum and maximum plausible levels, and increments between levels small enough to have a reasonable chance of finding the best treatment.
2. An easy-to-use and interpret plant rating scale.
3. A number of plants per rater that does not compromise the data by causing rater fatigue.

Considerations for the analysis. Horticultural experiments with objectively measured numeric responses usually use analysis of variance (ANOVA) or multiple regression or both. ANOVA and regression methods are based on the assumption that observations have a normal distribution and that a valid estimate of experimental error variance can be obtained from the error mean square. In quality assessment experiments, both of these assumptions may be violated. In order to obtain a valid analysis, alternatives to standard ANOVA and regression must be used.

The normality assumption is violated because the data are categories, not numbers measured on a continuous scale, like height or weight. For a two-point scale such as "like it" vs. "don't like it," the data have a binomial distribution. A binomial response uses the sample proportion of raters classifying the plant favorably. If the number of raters per plant is sufficiently large—ideally, 20 or more, but at least 10—the sample proportion is approximately normal and can be used directly in the analysis.

If a scale with three or more categories is used, the distribution is multinomial. In this case, the simplest approach is to assign nu-

Received for publication 9 Sept. 1996. Accepted for publication 5 Aug. 1997. Agricultural Research Division, Univ. of Nebraska Journal Series No. 11679. The cost of publishing this paper was defrayed in part by the payment of page charges. Under postal regulations, this paper therefore must be hereby marked *advertisement* solely to indicate this fact.

[1]To whom reprint requests should be addressed.

[2]Research done in partial fulfillment for the Master of Science degree.

meric codes to the categories that reflect their natural ordering from worst to best. For example, if the scale "unacceptable," "discount store quality," "florist quality" is used, the numeric code could be 0, 1, and 2, respectively, provided the equal spacing between the coded values reflects equal subjective distance between the categories. Alternatively, a code of 0, 2, and 3 may be preferred if "unacceptable" is regarded as more unlike "discount store quality" than the other two categories are from one another. Agresti (1990) discusses the use of numeric codes for ordered multinomial variables. Provided that the categories are well chosen, that raters can work with them, and that there is general agreement about the subjective distance among categories, acceptable results may be obtained from ANOVA and regression using the numeric codes as response variables. When general agreement about subjective distance does not exist, cumulative logit models can be used. Those models are beyond the scope of this paper, but Agresti (1990) discusses them in detail.

Estimating experimental error is difficult, because designs with little or no replication are usually required in order to maximize the number of treatment combinations but minimize the number of plants evaluated by each individual. As a result, experimental error generally has few or no degrees of freedom. Half-normal plots (Milliken and Johnson, 1989) are used to overcome this problem.

To summarize, the two main considerations in analysis are:

1. The data are neither continuous nor normal. For binary rating scales, this means the relevant response is the sample proportion. For multinomial rating scales, a numeric rating code that reflects the subjective distance among categories is used and these ratings are usually handled as continuous data. In either case, a sufficient number of raters per plant (ideally, 20 or more, but at least 10) must be used.

2. The estimate of experimental error will generally have few or no degrees of freedom. Alternative methods of analysis, like half-normal plots, are required.

Table 1. Example of assignment of treatments to groups of raters in incomplete blocks (Design 3).

N level:	1, 3, 5	1, 3, 5	2, 4, 6	2,4,6
S level:	1, 3, 5	2, 4, 6	1, 3, 5	2,4,6
Group 1	X			
Group 2		X		
Group 3			X	
Group 4				X

All combinations of N × S levels observed by group with X in given column. For example, Group 1 observes all combinations of levels 1,3, and 5 of N and levels 1, 3, and 5 of S. Each group has at least 10 raters.

Option 1: Each rater evaluates *n* plants per assigned N–S combination.

Option 2: Plants divided into *n* blocks, each block containing all 36 N–S combinations. Each block has four groups of raters, with treatment assignment as above. Thus, each rater evaluates one plant per assigned N–S combination.

METHODS OF ANALYSIS

Designs can address the need for a wide range of treatment levels at fairly small increments and the need to have each person evaluate sufficiently few plants so that "rater fatigue" does not compromise rating accuracy. The conflict implied by these two requirements can be addressed in a number of ways:

1. Include only one plant per treatment level or treatment level combination.

2. Have *n* plants per treatment level (or combination). Split the raters into *n* groups. Have each group rate one plant per treatment level (or combination).

3. Have *n* plants per treatment level (or combination). Have *m* groups or raters, where $m > n$. Assign raters to groups according to an incomplete factorial design.

Statistical designs that address these points are listed in order of increasing desirability. Researchers will generally find that this list coincides with increasing difficulty in implementation. Some compromise between statistical advantages and disadvantages, as listed below, and implementation ease is usually necessary.

Design 1. Include only one plant per treatment. Each of *r* different raters rate all plants. This is the simplest design to conduct and requires the fewest plants. However, this design allows for no replication of plants in the evaluation, resulting in no direct measure of experimental error. Only variation among raters, i.e., sampling unit variation, can be measured directly. If this design is used, indirect measures of experimental error, discussed in the next section, must be used.

Another disadvantage of this design is that the estimated effect of a given treatment is more likely to be distorted by a single unusual plant, although the number of treatment levels should alleviate the effect of an unusual plant. One can further minimize these effects by growing the plants in a replicated trial, then selecting a plant from each treatment. Selection should be random, to prevent bias, but obvious outliers, such as damaged plants, should be eliminated.

Design 2. Have n plants per treatment and n groups of raters. Statistically, this is preferable, because it provides a direct measure of experimental error, and minimizes the effects of unusual individual plants. Logistically, however, the larger number of plants required may be more difficult to manage. Also, more coordination is required to make sure that people rate their assigned plants and that the data are accurately recorded.

Design 3. Have n plants per treatment and m groups of raters. Assign raters to plants using an incomplete factorial. For example, suppose that the two treatment factors are levels of nitrogen (N) and sulfur (S) applied and that there are six levels of each. Two options for implementing this design are given in Table 1. Both options assign each rater to evaluate one plant for each of the nine treatment combinations. Statistically, this is preferable because it provides a direct measure of experimental error and substantially reduces the likelihood of rater fatigue. This design will almost certainly provide more accurate data, but is more difficult to conduct, and more raters are required than in the previous designs.

As mentioned previously, at least 10 but preferably 20 or more raters are required for a valid estimate of sample proportion for binomial ratings or a valid analysis of multinomial ratings. The three designs satisfy these requirements as follows. The first design requires at least 10 people rating each plant. The second design requires at least 10 raters per

Table 2. General form of the analysis of variance (ANOVA) table for Designs 1, 2, and 3.

Source	Degrees of freedom	Expected mean square[z]
Block	$n-1$	
A	$a-1$	$\sigma_R^2 + r\sigma_P^2 + \phi_A$
A_1	1	$\sigma_R^2 + r\sigma_P^2 + \phi_{A1}$
A_2	1	$\sigma_R^2 + r\sigma_P^2 + \phi_{A2}$
.		
.		
.		
A_{a-1}	1	$\sigma_R^2 + r\sigma_P^2 + \phi_{A,a-1}$
B	$b-1$	$\sigma_R^2 + r\sigma_P^2 + \phi_B$
B_1	1	$\sigma_R^2 + r\sigma_P^2 + \phi_{B1}$
B_2	1	$\sigma_R^2 + r\sigma_P^2 + \phi_{B2}$
.		
.		
.		
B_{b-1}	1	$\sigma_R^2 + r\sigma_P^2 + \phi_{B,b-1}$
A × B	$(a-1)(b-1)$	$\sigma_R^2 + r\sigma_P^2 + \phi_{AB}$
AB_1	1	$\sigma_R^2 + r\sigma_P^2 + \phi_{AB1}$
AB_2	1	$\sigma_R^2 + r\sigma_P^2 + \phi_{AB2}$
.		
.		
.		
$AB_{(a-1)(b-1)}$	1	$\sigma_R^2 + r\sigma_P^2 + \phi_{AB,(a-1)(b-1)}$
Exp. error	$(n-1)(t-1)$	$\sigma_R^2 + r\sigma_P^2$
Sampling error	$n(r-1)(t-1)$	σ_R^2

[z] σ_R^2 is the variance among raters, σ_P^2 is the variance among plants, and ϕ_{effect} depends on the sum of squares of the subscripted effect (A, B, AB, or contrast). For example, $\phi_A = [rn/(a-1)]\sum\alpha_i^2$, where α_i is the effect of the i^{th} level of A.

replication of plants, for a total of $10n$ raters. The third design requires at least 10 raters per incomplete factorial if the plants are not replicated, $10n$ if they are. Thus, the incomplete factorial design described above requires at least $40n$ raters.

Analysis. The basic form of the analysis for all three designs is a randomized block design with sub-sampling. Typically, treatments are factorial combinations, e.g., a levels of factor A and b levels of factor B. Also, partitioning sources of variation into single degree-of-freedom contrasts is generally desirable. Using the A × B factorial as a working example, the ANOVA for all three designs is given in Table 2.

The appropriate F-ratio to test treatment effects uses MS(exp. error) as the denominator. For example, the F-statistic to test a given treatment effect in the ANOVA (Table 2) is:

$$F = \frac{MS(treatment\ effect)}{MS(exp.\ err.)}$$

MS(treatment effect) is replaced by MS(A), MS(B), or MS(A × B) or MS(contrast), depending on the desired test.

Details of Design 1 analysis. As discussed in the previous section on design, consumer evaluation studies for which simplicity or minimum expense is a high priority will typically have only one plant per treatment combination, i.e., $n = 1$; therefore, the degrees of freedom for experimental error is 0, and a conventional F-test is not possible. If the unreplicated design is to be used, some alternative method of analysis is required.

One tempting, but grossly incorrect, alternative is to use the MS(sampling error) as the denominator term in the F-statistic. However, if this is done, one can see from the expected mean squares that:

$$F = \frac{MS(treatment\ effect)\ \ estimates}{MS(samp.\ err.)}$$

$$\frac{\sigma_R^2 + r\sigma_P^2 + \phi_{effect}}{\sigma_R^2}$$

This F-statistic may be large simply because σ_P^2, the variability among plants, is large, even if the treatment effect, ϕ_{effect}, is zero. The ratio of MS(treatment effect) to MS(sampling error) is a valid test of treatment effect only if one assumes $\sigma_P^2 = 0$, i.e., there is no random variation among plants. Clearly, this is an indefensible assumption.

The method of half-normal plots is appropriate for unreplicated designs. In the half-normal plot analysis, all treatment sums of squares are partitioned into complete sets of orthogonal contrasts. Each contrast sum of squares has the following properties:

1. One degree of freedom.
2. A χ^2 probability distribution.
3. An expected value of $\sigma_R^2 + r\sigma_P^2 + \phi_{contrast}$, where $\phi_{contrast}$ is the effect of that particular treatment contrast.

If a given contrast effect is zero, then its sum of squares has 1) an expected value equal to the expected value of the mean square for experimental error, $\sigma_R^2 + r\sigma_P^2$, and 2) a central χ^2 distribution. The one degree of freedom

Table 3. Analysis of variance of data on effects of N and S on quality of poinsettia plants with N and S effects partitioned into 1 d.f. contrasts; form of contrasts determined by *PROFILE ANALYSIS* technique.

Source	df	Sum of squares	Mean square	F value	$P > F$
Model	55	314.88080	5.72511	23.73	0.0001
Error	1231	297.04150	0.24130		
Corrected total	1286	611.92230			

Source[z, y]	df	Type I SS	Mean square	F value	$P > F$
N1	1	0.07764	0.07764	0.32	0.5707
N2	1	0.17495	0.17495	0.73	0.3947
N3	1	0.00207	0.00207	0.01	0.9262
N4	1	0.01118	0.01118	0.05	0.8296
N5	1	2.33669	2.33669	9.68	0.0019
N6	1	0.72262	0.72262	2.99	0.0838
N7	1	26.62339	26.62339	110.33	0.0001
S1	1	1.83696	1.83696	7.61	0.0059
S2	1	4.18841	4.18841	17.36	0.0001
S3	1	5.98958	5.98958	24.82	0.0001
S4	1	0.45679	0.45679	1.89	0.1691
S5	1	12.15236	12.15236	50.36	0.0001
S6	1	164.02383	164.02383	679.75	0.0001
NS1_1	1	0.04348	0.04348	0.18	0.6713
NS1_2	1	2.84058	2.84058	11.77	0.0006
NS1_3	1	0.35507	0.35507	1.47	0.2253
NS1_4	1	3.16957	3.16957	13.14	0.0003
NS1_5	1	0.00072	0.00072	0.00	0.9563
NS1_6	1	0.27381	0.27381	1.13	0.2870
NS2_1	1	6.39130	6.39130	26.49	0.0001
NS2_2	1	0.07729	0.07729	0.32	0.5715
NS2_3	1	0.11836	0.11836	0.49	0.4838
NS2_4	1	9.50870	9.50870	39.41	0.0001
NS2_5	1	8.08913	8.08913	33.52	0.0001
NS2_6	1	0.03882	0.03882	0.16	0.6884
NS3_1	1	5.88587	5.88587	24.39	0.0001
NS3_2	1	0.31944	0.31944	1.32	0.2501
NS3_3	1	0.03653	0.03653	0.15	0.6973
NS3_4	1	0.01467	0.01467	0.06	0.8053
NS3_5	1	0.05326	0.05326	0.22	0.6386
NS3_6	1	0.48525	0.48525	2.01	0.1564
NS4_1	1	9.00109	9.00109	37.30	0.0001
NS4_2	1	0.66993	0.66993	2.78	0.0959
NS4_3	1	0.50888	0.50888	2.11	0.1467
NS4_4	1	0.82272	0.82272	3.41	0.0651
NS4_5	1	4.78616	4.78616	19.83	0.0001
NS4_6	1	0.08701	0.08701	0.36	0.5483
NS5_1	1	0.14203	0.14203	0.59	0.4431
NS5_2	1	0.55652	0.55652	2.31	0.1291
NS5_3	1	4.04457	4.04457	16.76	0.0001
NS5_4	1	8.22964	8.22964	34.11	0.0001
NS5_5	1	0.84174	0.84174	3.49	0.0620
NS5_6	1	0.00253	0.00253	0.01	0.9185
NS6_1	1	0.40580	0.40580	1.68	0.1949
NS6_2	1	1.79503	1.79503	7.44	0.0065
NS6_3	1	8.55978	8.55978	35.47	0.0001
NS6_4	1	3.10688	3.10688	12.88	0.0003
NS6_5	1	5.24845	5.24845	21.75	0.0001
NS6_6	1	0.05557	0.05557	0.23	0.6314
NS7_1	1	0.68478	0.68478	2.84	0.0923
NS7_2	1	0.00207	0.00207	0.01	0.9262
NS7_3	1	2.66621	2.66621	11.05	0.0009
NS7_4	1	0.09321	0.09321	0.39	0.5344
NS7_5	1	4.37166	4.37166	18.12	0.0001
NS7_6	1	1.90022	1.90022	7.87	0.0051

[z]Main effect contrasts defined as follows:

N1: 250 vs. 275 ppm N
N2: 225 vs. mean of 250 and 275 ppm N
N3: 200 vs. mean of 225, 250, and 275 N
N4: 175 vs. mean of 200, 225, 250 and 275 N
N5: 150 vs. mean of 175, 200, 225, 250 and 275 N
N6: 125 vs. mean of 150, 175, 200, 225, 250 and 275 N
N7: 100 vs. mean of 125, 150, 175, 200, 225, 250 and 275 N
S1: 62.5 vs. 75 ppm S
S2: 50 vs. mean of 62.5 and 75 ppm S
S3: 37.5 vs. mean of 50, 62.5 and 75 ppm S
S4: 25 vs. mean of 37.5, 50, 62.5 and 75 ppm S
S5: 12.5 vs. mean of 25, 37.5, 50, 62.5 and 75 ppm S
S6: 0 vs. mean of 12.5, 25, 37.5, 50, 62.5 and 75 ppm S

[y]NS_i_j is interaction between i^{th} N contrast (N_i) and j^{th} S contrast (S_j), e.g., NS_1_1 is N1 × S1.

central χ^2-distribution arises, by definition, as the square of a standard normal random variable. Thus, the square roots of the contrast sum of squares whose corresponding treatment effects are zero would be expected to behave like standard normal random variables.

In the half-normal plot analysis, the square roots of all the contrast sums of squares are plotted. Since the square root can be positive or negative, but the signs of the square roots of the contrast sums of squares are unknown, their absolute values are used in practice. The half-normal plot is a plot of the absolute values of standard normal random variables. If the X- and Y-axes are scaled properly, standard normal random variables will form a straight line. Thus, values for all contrasts whose effects are zero will fall on an approximately straight line, whereas those whose effects are non-zero will tend to deviate from the straight line. Thus, contrasts that deviate from the straight line may be statistically significant. Their formal test is described below.

The contrasts that fall on the line can be used to estimate experimental error; since they behave like normal random variables, we can infer that their component of treatment effect is zero. Thus, each such contrast has an expected value of $\sigma_R^2 + r\sigma_P^2$. Pooling all such contrasts yields a reasonable alternative to the MS(experimental error). The estimate of experimental error can be used to formally test the contrasts whose effects were suspected to be significant by their deviation from straight line.

Analysis of Design 2. Design 2 is a randomized complete-block design with sub-sampling. The ANOVA given above (Table 2) can be used without modification. This is because the number of plants per treatment (n) is greater than one, and hence there are $(n-1)(t-1) > 0$ degrees of freedom for error.

Analysis of Design 3. Design 3 (Table 1) is an incomplete block design with sub-sampling, where the incomplete blocks correspond to the groups of raters. The same form of the ANOVA given above (Table 2) is valid. The degrees of freedom for experimental error are greater than zero, hence standard F-tests for treatment effects may be used. However, the specific degrees of freedom and coefficients for the expected mean squares vary with the design. For example, the degrees of freedom for block is $m-1$, determined by the number of rater groups, rather than $n-1$ as in the ANOVA tables for Designs 1 and 2. These are generally handled correctly by appropriate computing software using the same model as for Design 2, e.g., SAS PROC GLM using Type III sums of squares. Most experimental design texts, e.g., Milliken and Johnson (1984), present the analysis of incomplete block designs.

SAMPLE ANALYSIS—QUALITY EVALUATION ON POINSETTIAS GROWN AT DIFFERENT LEVELS OF N AND S

An experiment to measure the effects of seven levels of sulfur (S) (0, 12.5, 25, 37.5, 50, 62.5, and 75 mg·L^{-1}) and eight levels of nitrogen (N) (100, 125, 150, 175, 200, 225, 250, and 275 mg·L^{-1}) applied to poinsettia (*Euphorbia pulcherrima* L.) plants was conducted at the Univ. of Nebraska–Lincoln. The experiment involved various physiological measurements, which were obtained using a randomized complete-block design with three blocks. At the end of the experiment, consumer and producer evaluations were also desired.

The design used for the producer evaluation was the unreplicated plant design described above. Twenty-three producers rated each plant using the following scale: 0 = plant not acceptable for commercial sale; 1 = plant acceptable, "discount store quality"; and 2 = fine florist quality plant.

Each producer rated 56 plants, one plant for each N × S treatment combination. The plant selected for each N × S combination was representative of the three replicates. Admittedly, by the criteria for rater fatigue discussed previously, each rater should have been asked to rate fewer plants. The main purpose of this section is to demonstrate the half-normal plot

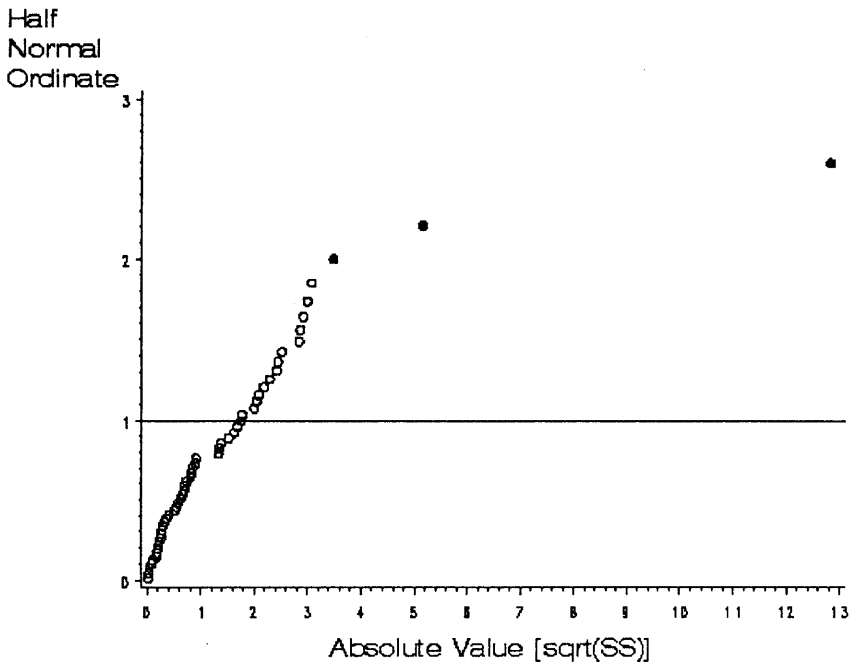

Fig. 1. Half-normal plot of ALL sums of squares in model, based upon ANOVA in Table 3.

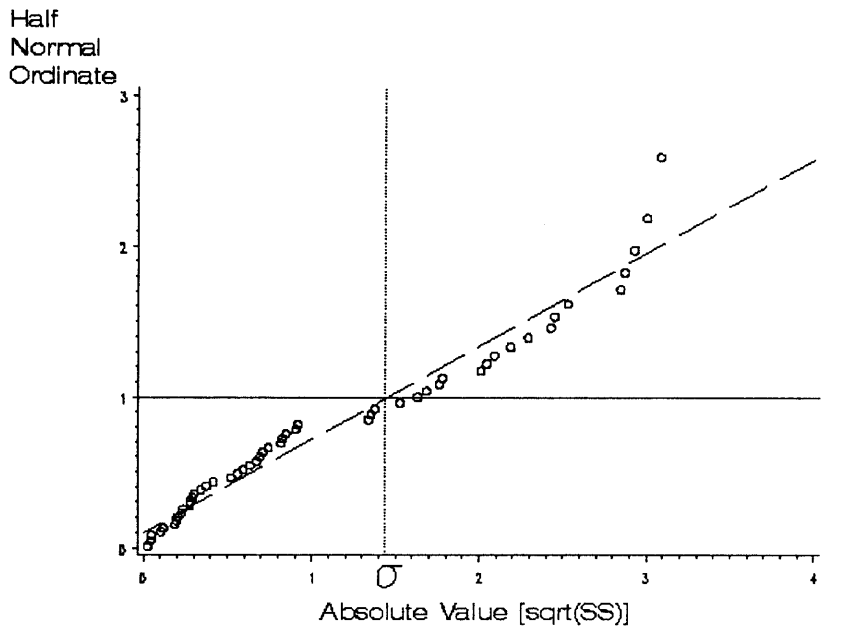

Fig. 2. Half-normal plot of remaining SS after deleting N7 (lowest N vs. others) and S6 (lowest S vs. others), and S5 (2 lowest S vs. others) effects, based on ANOVA of Table 3.

analysis. In the interest of clarity, we use the simplest possible design, despite its drawbacks.

The basic ANOVA was:

Source	df
N	7
S	6
N × S	42
exp. error	0
samp. error	1232

One rating was illegible and the observation was thus discarded, leaving a total of 1287 valid observations and 1231 degrees of freedom for sampling error. However, there are zero degrees of freedom for experimental error, and thus no standard F-ratios for N, S, or N × S.

The half-normal plot method was used to estimate experimental error. The N and S main effects were partitioned into one degree of freedom contrasts using profile analysis contrasts. That is, the mean rating for the highest level was contrasted to the mean for the next highest level, then the average of the two highest levels was contrasted to the third highest level, etc. For example, the specific set of contrasts used for S were

S_1: $\mu_{75} - \mu_{62.5}$
S_2: $\mu_{75} + \mu_{62.5} - 2\mu_{50}$
S_3: $\mu_{75} + \mu_{62.5} + \mu_{50} - 3\mu_{37.5}$
S_4: $\mu_{75} + \mu_{62.5} + \mu_{50} + \mu_{37.5} - 4\mu_{25}$
S_5: $\mu_{75} + \mu_{62.5} + \mu_{50} + \mu_{37.5} + \mu_{25} - 5\mu_{12.5}$
S_6: $\mu_{75} + \mu_{62.5} + \mu_{50} + \mu_{37.5} + \mu_{25} + \mu_{12.5} - 6\mu_0$

where μ_i is the mean of the i^{th} level of sulfur applied. The profile analysis is a particularly useful set of contrasts for identifying the minimum necessary treatment. When the contrasts are tested in order, starting with S_1, then S_2, etc., the optimum treatment is between the two lowest levels of first significant contrast. For example, suppose S_1 and S_2 are not significant, but S_3 is. Then there is no evidence to conclude that mean acceptance ratings differ among the three highest S levels, 50, 62.5, and 75 mg·L^{-1}, but there is evidence to conclude that the mean rating for sulfur level 37.5 mg·L^{-1} is significantly less than those for the higher treatment levels. Thus, at least 37.5 but <50 mg·L^{-1} S is required to achieve maximum acceptance.

The main effect of N was partitioned using a set of profile contrasts analogous to those constructed for S. The N × S interaction was partitioned by constructing an interaction contrast for every combination of the N and S main effect profile analysis contrasts.

The ANOVA is given in Table 3. Note that two contrasts, S6 and N7, have noticeably larger sums of squares than all the other contrasts, and another, S5, has a sum of squares that is greater than the rest, though not as great as S6 and N7. Also, the F-values computed are based on MS(sampling error) in the absence of an estimate of experimental error. Most of these F-values have p-values that appear to be significant. Recalling that F-values computed in this way may be significant either because there is a treatment effect or because there is random variation among plants, it seems reasonable to speculate that S6 and N7, and possibly S5, reflect real treatment effects, whereas all the other contrasts are merely detecting random variation among plants. Constructing half-normal plots addresses this conjecture formally.

Table 4. Analysis of variance of effects of N and S on quality of poinsettia plants with N7, S5, and S6 in model and all other model effects pooled to estimate experimental error. Error term labeled N*S.

Source	df	Sum of squares	Mean square	F value	$P > F$
Model	55	314.88080	5.72511	23.73	0.0001
Error	1231	297.04150	0.24130		
Corrected total	1286	611.92230			
	R-Square		CV	Root MSE	RATING mean
	0.514576		48.74366	0.4912	1.0078

Source	df	Type I SS	Mean square	F value	$P > F$
S5	1	12.15236	12.15236	50.36	0.0001
S6	1	163.94031	163.94031	679.40	0.0001
N7	1	26.53904	26.53904	109.98	0.0001
N*S (error)	52	112.24909	2.15864	8.95	0.0001

Tests of hypotheses using the Type I MS for N*S as an error term

Source	df	Type I SS	Mean square	F value	$P > F$
S5	1	12.15236	12.15236	5.63	0.0214
S6	1	163.94031	163.94031	75.95	0.0001
N7	1	26.53904	26.53904	12.29	0.0009

Table 5. Alternative (preferred) analysis of variance of effects of N and S on quality of poinsettia plants—N and S main effects in model, MS(N*S) used to estimate experimental error, σ^2.

Source	df	Sum of squares	Mean square	F value	$P > F$
Model	55	314.88080	5.72511	23.73	0.0001
Error	1231	297.04150	0.24130		
Corrected total	1286	611.92230			

Source	df	Type I SS	Mean square	F value	$P > F$
N1	1	0.07764	0.07764	0.32	0.5707
N2	1	0.17495	0.17495	0.73	0.3947
N3	1	0.00207	0.00207	0.01	0.9262
N4	1	0.01118	0.01118	0.05	0.8296
N5	1	2.33669	2.33669	9.68	0.0019
N6	1	0.72262	0.72262	2.99	0.0838
N7	1	26.62339	26.62339	110.33	0.0001
S1	1	1.83696	1.83696	7.61	0.0059
S2	1	4.18841	4.18841	17.36	0.0001
S3	1	5.98958	5.98958	24.82	0.0001
S4	1	0.45679	0.45679	1.89	0.1691
S5	1	12.15236	12.15236	50.36	0.0001
S6	1	164.02383	164.02383	679.75	0.0001
N*S (error)	42	96.28433	2.29248	9.50	0.0001

Tests of hypotheses using the Type I MS for N*S as an error term

Source	df	Type I SS	Mean square	F Value	$P > F$
N1	1	0.07764	0.07764	0.03	0.8549
N2	1	0.17495	0.17495	0.08	0.7837
N3	1	0.00207	0.00207	0.00	0.9762
N4	1	0.01118	0.01118	0.00	0.9447
N5	1	2.33669	2.33669	1.02	0.3185
N6	1	0.72262	0.72262	0.32	0.5775
N7	1	26.62339	26.62339	11.61	0.0015
S1	1	1.83696	1.83696	0.80	0.3758
S2	1	4.18841	4.18841	1.83	0.1837
S3	1	5.98958	5.98958	2.61	0.1135
S4	1	0.45679	0.45679	0.20	0.6576
S5	1	12.15236	12.15236	5.30	0.0263
S6	1	164.02383	164.02383	71.55	0.0001

The half-normal plot of all contrast sums of squares (Fig. 1) reveals that all square roots of sums of squares "line up" on a straight line with three exceptions—two obvious, the other more subtle. The S6 and N7 contrasts, and arguably S5 as well, deviate below and well to the right of the line. From this graph, we can conclude that all of the contrasts except S5, S6, and N7 are simply measuring random variation among plants. Contrast S5 also may be measuring no more than random variation, but for the sake of caution it should be provisionally treated as if significant.

Experimental error can be measured in one of three ways. One is to make a new half-normal plot using only the nonsignificant sums of squares—in this case, all contrasts except

S5, S6, and N7—and fit a straight line (Fig. 2). The point on the X-axis corresponding to where the line crosses Y = 1 is an estimate of σ. By inspection of Fig. 2, σ is $\approx$1.45. Thus, its square, 2.10, is an estimate of σ^2, and is used in place of MS(experimental error).

A second way to estimate experimental error is to pool all 52 nonsignificant contrasts into one mean square (Table 4). Conceptually, this is similar to the previous approach, but the estimate of σ^2 does not depend on approximation from a graph. The mean square labeled "N*S" in Table 4 is actually the mean square for the pooled nonsignificant contrasts. Thus, it estimates σ^2 and is used in the subsequent analysis in place of MS(experimental error) to calculate F-values for S5, S6, and N7. Using this method, MS(experimental error) is 2.16, whereas it was 2.10 using the approximation from Fig. 2.

A third way, preferable when justified by the data, is to use the highest order interaction mean square as the estimate of experimental error. This can be done legitimately if 1) all contrasts forming the partition of the interaction are nonsignificant according to the half-normal plot, and 2) the interaction has sufficient degrees of freedom for a valid F-test. Statisticians disagree on exactly how many degrees of freedom are "sufficient" but 15–20 is generally considered adequate. In this example, the N × S interaction clearly meets both of these requirements. In the resulting ANOVA (Table 5), the S5, S6, and N7 contrasts are all significant ($P < 0.05$). One can conclude from these data that the minimum N level required to achieve maximum acceptance is between 100 and 125 mg·L^{-1} and the minimum S level is between 12.5 and 25 mg·L^{-1}. Observe that treating S5 cautiously was justified. When in doubt, treat an effect as potentially significant.

The third method is preferable for two reasons. First, the treatment effects associated with some of the nonsignificant contrasts may actually exist; that is, some nonsignificant contrasts may represent type two errors. These type two errors are least likely to result in serious violations of the assumptions of analysis of variance if the mean square for experimental error involves only the highest order interaction. The second reason to prefer using high-order interactions to estimate experimental error is that incomplete factorial designs can be used in future quality assessments. While this has no direct bearing on the analysis, it is extremely useful for planning future experiments.

SUMMARY AND CONCLUSIONS

Plant quality assessment experiments in ornamental horticulture have several specific considerations that must be addressed in their design and analysis. The response variable is a subjective rating. The design must provide raters with an easy-to-use rating scale and allow for their tendency to become fatigued. At the same time, a wide range of treatment levels must be observed, and the increments between levels must be relatively small. For these reasons, unreplicated designs, or, when justified by experience, incomplete factorial designs are essential.

Both unreplicated and incomplete factorial designs present problems in obtaining direct estimates of experimental error. Indirect methods, such as half-normal plots, must be used for this purpose. The half-normal plot method is particularly desirable because it is easy to perform and to interpret.

The methods presented in this paper provide horticulturists with statistically valid procedures for measuring treatment effects on producer and consumer acceptability of ornamental plants, and may be used by researchers and plant producers as well. For example, a producer contemplating a change in some management procedure could obtain data regarding the potential effect of the change on consumer acceptance before proceeding. Also, these methods have been used with great success in a variety of quality improvement contexts. Based on our experience at the Univ. of Nebraska, we suggest that other horticulturists may find these methods, particularly half-normal plots, useful for evaluating plant quality.

Literature Cited

Agresti, A. 1990. Categorical data analysis. Wiley, New York.

Box, G.E.P., W.G. Hunter, and J.S. Hunter. 1978. Statistics for experimenters. Wiley, New York.

Cloninger, M.R., R.F. Baldwin, and G.F. Krause. 1976. Analysis of sensory rating scales. J. Food Sci. 41:1225–1228.

Cox, E.P., III. 1980. The optimal number of response alternatives for a scale: A review. J. Market Res. 17:407–422.

Freeman, D.H. 1987. Applied categorical data analysis. Marcel Dekker, New York.

Mead, R. 1988. The design of experiments. Cambridge Univ. Press, New York.

Milliken, G.A. and D.E. Johnson. Analysis of messy data. vol. 1. Designed experiments. Van Nostrand Reinhold, New York.

Milliken, G.A. and D.E. Johnson. 1989. Analysis of messy data. vol. 2. Nonreplicated experiments. Van Nostrand Reinhold, New York.

SAS Institute, Inc. 1989. SAS user's guide. vers. 6. 4th ed. SAS Inst., Cary, N.C.

Stone, H. and J.L. Siedel. 1993. Sensory evaluation practices. 2nd ed. Academic, San Diego.