Data Collection and Statistical Topics for the Preparation and Review of Manuscripts

Jason Osborne[1] and Eric Simonne[2]

Additional index words. cultivar, cultigen, variety recommendation, mean comparison procedures, inference

Summary. The challenges encountered and discussions generated during the review process of the manuscripts submitted to the Variety Trials category of HortTechnology have revealed the need to review issues encountered during manuscript preparation and to provide flexible guidelines for authors and reviewers. Using a question/answer format, this manuscript discusses issues related to data collection and statistical methods available to compare varieties. Clear objectives and conclusions, adequate plot size, careful selection of entries, and sound statistical procedures are considered essential. Several additional factors (following standard production practices, using multiple seed sources, reporting analysis of variance table and mean square error, reporting multiyear/multilocation trials) are regarded as desirable, with different degrees of desirability, depending on the crop. These flexible guidelines should be viewed as recommendations for authors and reviewers rather than requirements. While defining the state-of-the-art in variety trialing is of interest to all those involved, it may be difficult to achieve when resources are limiting. It is ultimately the prerogative and responsibility of the author(s) to ensure that the work is scientifically sound.

In the author instructions included in every issue of *HortTechnology*, the category Variety Trials is defined as the repository of "articles reporting the results of studies in which varieties or species are evaluated for comparative performance. Manuscripts should be oriented toward testing and differentiating varieties using traits of interest to growers, industry representatives, and other professional horticulturists." Manuscripts submitted to the Variety Trials category of *HortTechnology* are peer-reviewed by at least three colleagues (including the associate editor) to "assure readers that the published papers have been found acceptable by competent, independent reviewers." As part of a refereed publication, variety trial manuscripts are expected to follow the criteria of excellence set forth by ASHS. The strength of refereed publications comes from 1) the trust in the quality of the work described and in the relevance of the methodology used, and 2) the scope of the inference that can be made from the experimental results. When applied to variety testing, these two concepts become 1) issues related to data collection, and 2) statistical methods available to compare varieties.

Journal paper R-08724 of the Florida Agricultural Experiment Station.

[1] Campus Box 8203, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203.

[2] Horticultural Sciences Department, 1241 Fifield Hall, PO Box 110690, University of Florida, Gainsville, FL 32611-0690.

Table 1. Essential and desirable traits of manuscripts submitted to the Variety Trials category of *HortTechnology*.

Trait	Essential	Highly desirable	Very desirable	Desirable
Stating objectives clearly	x			
Reporting standardized quality rating for the trials				x
Defining and discussing the area where trial results may apply				x
Following standard production practices		x		
Including reference variety(s) in trial		x		
Using multiple seed sources			x	
Using replicated data		x		
Basing conclusions on multi-year/location		x		
Including factor used to convert unit/plot to unit/acre		x		
Reporting multiple attributes		x		
Using adequate plot size	x			
Justification of entry selection			x	
Using global indices			x	
Including power calculation			x	
Reporting analysis of variance (ANOVA) and mean square error (MSE)	x			
Inspecting and reporting status of ANOVA assumptions	x			
Using nonparametric statistics when appropriate	x			
Ending manuscript with a clear conclusion	x			

Traditionally, results of variety trials have been published in nonrefereed publications such as bulletins, station reports, regional reports, newsletters, or other clientele-oriented publications. The articles that have been published in the Variety Trials category of *HortTechnology* as of 2001 illustrate the similarities and the differences in publishing applied research and publishing variety trial results. The review process itself revealed challenges in reaching a consensus among authors, reviewers, and associate editors on what constitutes an acceptable manuscript. It is clear that all parties involved need to be able to refer to guidelines describing what an acceptable variety trial manuscript should be like. The objective of this article is to explore some essential and desirable characteristics of manuscripts submitted to the Variety Trials category of *HortTechnology* (Table 1). This paper is intended to serve as a starting point for authors, reviewers and associate editors, not as a set of requirements. Hopefully, those who feel the need to modify, expand, challenge, or update this list will feel stimulated to develop

Table 2. Statement of objective, conclusion of selected articles published (1998–2001) in the Variety Trials category of *HortTechnology*.

Statement of objective
We evaluated 11 varieties of trachelium for vegetative and flowering characteristics when produced in ground beds as cut flowers.
The objective of this study was to determine if hybrid varieties outperform OP varieties in terms of plant and bulb traits when grown in southern New Mexico.
...field trials were conducted to evaluate plant, yield, and pod characteristics of filet snap bean variety.
This study objective was to determine the variation between taro varieties in the quality of corms used to make a boiled or microwaved food, or made into poi and fried as chips.
The objective of this study was to focus specifically on full sun perennials.... Results from this study will assist horticultural professionals and consumers in similar climates with selection of flowering perennials.
[none worded specifically]
The objectives of this study were to 1) evaluate field performance, ear characteristics, and eating quality of selected white sweet corn varieties, 2) globally compare varieties using an overall rank-sum index, and 3) determine if ‘Silver Queen’ is still the best variety or if it benefits from name recognition, or both.
We have evaluated the horticultural characteristics of 10 new selections of ‘Mariana 2616’...
The objective of these trials was to evaluate field performance in terms of yield, quality, and virus resistance of current cultigens supplied by various seed companies.

[z] Published in the Research Updates category of *HortTechnology*.

the missing research data, or provide in-depth summaries of existing literature. Our general objective was to propose uniform quality standards for manuscripts submitted to the Variety Trials category of *HortTechnology*. For clarity of presentation and because topics discussed here have been controversial, we used the question/answer format. Questions are divided into two groups: those addressing data collection issues, and those addressing statistical issues.

Questions related to methodology and data collection

Question 1. Should the word **cultivar** or **variety** be used? ASHS currently regards cultivar or variety as synonymous. It is requested that authors choose one term and use it throughout the manuscript to the exclusion of the other. The use of the term **cultigen** to describe a group of entries made up of varieties, advanced breeding lines, and/or clones is not considered appropriate for an outreach publication like *HortTechnology*.

Question 2. What are the goals of variety evaluation? Variety trials are used by horticulturists to develop and update variety recommendations, by growers to identify superior varieties, and by breeders to identify the poorest varieties. The choice of the goal has consequences on the experimental plan and statistical analysis (Wehner, 1987). The quest for the superior variety may involve different techniques, depending on whether a reference variety (industry standard) already exists or not (Eskridge and Mumm, 1992). While detecting the superior variety requires an inference (and therefore replicated data; Crossa, 1990; Fernandez, 1991), eliminating poor varieties may be done based on observation only. That is, statistical inference may not be needed to eliminate a variety that exhibits some disqualifying attributes, such as unacceptable shape, without a variance estimate. Rapid evaluation of breeding lines, clones and populations has been done by breeders and others for years. Recently, a program called Rapid Action Cultivar Evaluation (RACE) was implemented at the University of Kentucky to identify poor breeding material and varieties from cooperative trials (Rowell, 2000).

Question 3. How should the objectives of variety testing be worded? The strength of a research manuscript comes in part from the coherence between the title, the abstract, the statement of objective, and the conclusion. The message conveyed by all these sections should be the same. In particular, a manuscript should have a clear goal. Possible goals for variety trials are "identify poor-performing varieties," "select superior (or best-performing) varieties," or "update recommendations" by comparing the performance of new varieties and/or advanced breeding lines to those of the industry standards. For example, articles so far published in the Variety Trials of category of *HortTechnology* that have specific goals also tend to have specific conclusions and a clear take-home message (Table 2). A clear take-home message is important to a large segment of the *HortTechnology* readership. These goals are clear because they will allow a specific message in the conclusion (Specific here refers to a clearly worded conclusion, whether the goal has been reached or not). The work will, therefore, make a significant contribution to the literature. Also, clear goals contain a blueprint of a statistical model since they suggest a comparison of means. However, goals such as "evaluate varieties," "grow new genotypes," or "try new plant materials" are not action goals, and they do not suggest a comparison of means. From past experience, manuscripts

Conclusion	Reference
[Selection of the best varieties]	Liang and Harbaugh, 2001
For onion production in southern New Mexico, OP varieties generally performed as well as, if not better, than hybrid varieties.	Cramer, 2001
The small sieve filet-type varieties were well adapted to mechanical harvest.	Mullins and Straw, 2001
The variation in taro and corm characteristics and differences in human preferences for fried chips, microwaved corm and poi made from different varieties made this widely distributed crop, adaptable to many situations.	Paull et al., 2000
[none worded specifically]	Kessler et al., 2000
Our trial results suggest that for Pennsylvania, no single variety will perform best during the entire growing season, and that varieties should be chosen based on a seasonal performance.	Orzolek et al., 2000
[one specific to each objective]	Simonne et al., 1999
Such a rootstock currently does not exist and none was discovered among those tested in these experiments.	Southwick et al., 1999
[none worded specifically]	Schultheis and Walters, 1998^{z}

Table 3. Descriptions of proposed standardized ratings for use in reporting variety trials.

Rating	Weather	Fertilization	Irrigation	Pest incidence	Overall growing conditions
5	Very good	Very good	Very good	None	Excellent
4	Favorable	Good	Good	Light	Good
3	Acceptable	Acceptable	Acceptable	Tolerable	Acceptable
2	Adverse	Low	Low	Adverse	Questionable
1	Destructive	Very low	Insufficient	Destructive	Useless

without a clearly stated goal seldom generate specific information. The reader is left wondering So what?

QUESTION 4. Should we include quality ratings for the trials? Several variety trial reports (Simonne, 1999b, 2000) have standardized the description of growing conditions of variety trials, and proposed ratings of weather conditions, fertilization, irrigation, pest incidence, and overall growing conditions (Table 3). This approach was adopted by several authors contributing the Southeastern Regional Bulletins who found these ratings practical and simple. These ratings may also be used as quality control when, for example, trial results with at least one rating below 3 (see Table 3 for corresponding description) are not reported.

When preparing a manuscript for *HortTechnology*, the use of quality ratings should be considered as an addition to the materials and methods section, and not a substitute. It is still essential to describe the fertilization and irrigation programs used during the trial, as they may influence the results or help explain year-to-year or location-to-location differences. Typically, variety trials should be conducted following current recommendations or industry practices. However, in situations where cultural practices used do not follow current recommendations, the practices used should be described and justified. The description should be detailed enough as to enable *HortTechnology* readers to reproduce the trial with similar results.

QUESTION 5. What geographical area do trial results apply to? Traditionally, variety recommendations are made for each state. With decreasing resources allocated to variety testing, multi-state or regional variety trial programs have emerged such as the ones in the southeastern U.S. (Simonne, 2000) or in the midwestern U.S. (Morales and Maynard, 2000). In some instances, regional vegetable produc-

Table 4. Guidelines for plot size and data collection used in the variety trial program in Florida (adapted from Maynard, 1987).

Crop	Plot size[z]	No. of harvests	Yield unit[z]
Bean, snap (*Phaseolus vulgaris*)	10–20 row ft	1–3	30-lb bushel
Broccoli (*Brassica oleracea* var *italica*) and cauliflower (*B. oleracea* var *botritys*)	20–40 plants	2–4	23-lb carton
Cabbage (*Brassica oleracea* var *capitata*)	20–40 plants	1–3	50-lb carton
Carrot (*Daucus carota*)	2-row bed, 20 ft long, the center 8–10 ft can be harvested or sections can be harvested at weekly intervals to determine optimum pack out	1 for each variety	50-lb unit

tion recommendations (including varieties) have been compiled (Sanders, 1999), but they are still presented on a state-by-state basis. Similarly, most articles currently published in the Variety Trials category of *HortTechnology* include a state name in the title. By encouraging authors to identify the geographical zone similar to that of the evaluation, progress could be made toward regional recommendations based on growing conditions (such as soil, climate, or production system) rather than state lines. Instead of describing a geographical area, authors could include a summary of relevant weather data (maximum/minimum temperature, chilling hours, rainfall) and soil type. This allows the reader to make informed choices on how similar his/her production conditions are to that of the test. Perhaps a map could be included showing the "area of potential application of the results presented in this work based on soil type, weather conditions, or planting seasons." Without reference to geographical area, inference on variety performance can seldom be legitimately extended beyond the region in which the trial was conducted.

Question 6. How many entries should be in a trial? The answer to this fundamental Question is not simple. Variety trials should contain a reference variety (current industry standards) together with new varieties and/or advanced breeding lines. At least one reference variety should be chosen when no industry standards are available. If only one reference variety is used, comparisons for the whole trial may be affected when it performs in an unusual fashion. Thus, it is preferable to include two or three checks. The two main factors that determine the number of entries are crop type and resources needed to perform the trial. For some crops such as watermelon (*Citrullus lanatus*), sweet corn (*Zea mays*), or tomato (*Lycopersicon esculentum*), many varieties are introduced each year (Simonne et al., 2000). For these crops, variety trials may include up to 30 entries. In contrast, few new releases occur each year for crops such as strawberry (*Fragaria* ×*ananassa*). For these crops, it is possible to have a valid variety trial with only two entries - the industry standard and the new introduction. The other factor determining the number of entries in a trial are resources—mainly in terms of space, labor, and cost—needed to perform the trial. For example, hand-harvesting takes an estimated 150 h/acre (371 h·ha^{-1}) for cucumber (*Cucumis sativus*) or eggplant (*Solanum melongena*; multiple harvests), but only 30 and 25 h/acre (74 and 62 h·ha^{-1}) for sweet corn, and watermelon if once-over harvested, respectively. For comparison, it only takes 20 h/acre (50 h·ha^{-1}) to harvest potato (*Solanum tuberosum*) mechanically (Brown et al., 1983).

Increasing the number of entries without increasing the number of replications will affect the statistical power, or ability to detect real differences among varieties in the experiment. While adding very different entries to a study can increase the statistical power to detect difference among all varieties, there is also the danger of losing power by adding similar entries. To see this, consider a completely randomized design of 10 total observations

Data acquisition	Observations
Days to first harvest	Susceptibility to disease
Early and total yield	
Plant height	
Pod height	
Pod shape: round, oval, or flat	
Pod straightness: qualitative	
Pod color: qualitative	
Pod removal force requirement: for machine harvest	
Days to first harvest	Susceptibility to disease and disorders such as hollow stem
Early and total yield	
Head or curd diameter	
Head or curd weight	Uniformity
Number of harvests	Head tightness
Percent marketable	Head cover
Head leafiness	
Head color	
Broccoli bead characteristics	
Days to first harvest	Susceptibility to disease
Yield expressed	Incidence of tipburn
Average head weight	
Core length	
Head tightness	
Plant color	
Head shape	
Percent marketable	
Days to harvest	Susceptibility to disease
Yield	Top vigor
Percent marketable	Top height
Root length	
Root diameter	
Root shape	
Root color	

Table 4. Guidelines for plot size and data collection used in the variety trial program in Florida (adapted from Maynard, 1987).

Crop	Plot size[z]	No. of harvests	Yield unit[z]
Celery (*Apium graveolens* var. *dulce*)	1 row, 20 ft long	1	60-lb carton
Sweet corn (*Zea mays*)	25 ft long, 1 row if varieties are of similar maturity, 3 rows for pollination if varieties differ widely in maturity. Endosperm types should be separated at least 500 ft, if possible.	1 for each variety	42-lb carton
Cucumber (*Cucumis sativus*)	12–25 plants	8–12	55-lb bushel
Lettuce (*Lactuca sativa*)	1 row, 20 ft long	1	50-lb carton
Muskmelon (*Cucumis melo*)	12–25 plants	6–12	cwt
Okra (*Abelmoschus esculentus*)	30–40 plants	20–30	30-lb bushel
Onion (*Allium cepa*)	2 rows, 10–20 ft long	1 for each variety	50-lb bag
Pepper (*Capsicum annuum*)	20–30 plants	3–5	25-lb bushel
Potato (*Solanum tuberosum*)	1 row, 20 ft long	1 for each variety	cwt
Radish (*Raphanus sativus*)	3–4 ft of row, 36–48 plants	1 per variety (varieties may require different harvest days)	12-lb carton

Continued

Data acquisition	Observations
Days to harvest	Susceptibility to disease
Yield	Incidence of defects such as pithiness, brown stem, nodal cracking, bolting
Percent marketable	
Stalk size, number per crate	
Petiole length and width	
Days to mid-silk to estimate harvest date	Susceptibility to disease
Yield expressed as 42-lb crates containing 4.5–5 dozen ears	Lodging
Number of marketable ears	Ease of snapping
Days to harvest	Flag leaves
Plant height	Kernel color, sweetness, tenderness
Ear length and diameter	Tendency for double ears
Husked ear length	
Number of kernel rows	
Tip fill	
Husk cover	
Days to first harvest	Susceptibility to disease
Early and total yield expressed as 55-lb bushels for pickling.	Cucumber Improvement Committee (PCIC) values for pickles
Fruit length and diameter	
Fruit color	
Yield	Susceptibility to disease
Days to harvest	Incidence of defects such as cracked rib, tipburn, and bolting
Head weight and firmness	Ability to hold in the field
Percent marketable	
Days to first harvest	Susceptibility to disease
Early and total yield	Fruit flesh color
Fruit weight	Presence of sutures
Fruit shape	Netting characteristics
Cavity dimensions	
Flesh width	
Soluble solids	
Early and total yields	Susceptibility to pests
Days to first harvest	
Pod color	
Pod shape	
Plant height	
Yields	Susceptibility to disease
Days to harvest	Incidence of defects such as bolting, rots
Percent marketable	Curing characteristics
Bulb diameter	
Neck diameter	
Bulb weight	
Early and total yields	Susceptibility to disease
Days to harvest	
Fruit weight	
Number of fruit per bushel	
Number of lobes	
Wall thickness	
Length/diameter ratio	
Yield	Susceptibility to disease
Days to harvest	Occurrence of tuber defects
Tuber shape	
Skin color	
Skin type (smooth, russet)	
Specific gravity	
Days to harvest	Susceptibility to disease
Average root weight	Top characteristics
Proportion of roots in size classes	
Marketable yield	
Proportion of roots that are marketable	
Incidence of defects, splits, cracks, misshapen, pithiness, black root rot	

Table 4. Guidelines for plot size and data collection used in the variety trial program in Florida (adapted from Maynard, 1987).

Crop	Plot size[z]	No. of harvests	Yield unit[z]
Squash (*Cucurbita pepo*) and pumpkin (*C. pepo, C. maxima, C. moschata*)	12–25 plants	summer squash = 12–20, pumpkin and winter squash = 1–3	42-lb bushel
Susceptibility to disease			
Strawberry (*Fragaria ×ananassa*)	12–20 plants	40–60	10-lb flat
Tomato (*Lycopersicum esculentum*)	10 plants	2–4	25-lb carton
Watermelon (*Citrullus lanatus*)	10–12 plants	3	cwt

[z] 1 ft = 0.3 m, 1 lb = 0.454 kg, 1cwt = 100 lb.

from five varieties, with ordered means (2, 4, 6, 6, 7) and SD = 1. The power of the F test with level $\alpha = 0.05$ can be shown to be 0.72. The power of the F test for the same experiment on 8 varieties with means (2, 4, 5, 5, 5, 6, 6, 7) decreases to 0.62. While the larger experiment has more information for estimation of the experimental error, thereby increasing the power of the test, the size is decreased and the number of parameters that have to be estimated from the data is increased, thereby decreasing the power. Simple SAS (1999) code using the {probf} and {finv} commands to perform these kinds of computations appears in the subsequent discussion of power. The exact nature of this effect can be investigated using power computations and software described below.

QUESTION 7. What are the typical plot sizes used in variety evaluation? Guidelines have been proposed (Maynard, 1987) regarding plot size and number of plants per plot for vegetable crops grown in Florida (Table 4). While these guidelines have been largely adopted nation-wide in variety evaluation, the coefficients of variation (CV) observed for total marketable yield were often above the 20% threshold (Table 5). In comparison, CV of small grains trials are usually below 5%. In general, CV for vegetable crops were the lowest for crops with high plant populations in each plot [carrot (*Daucus carota*), turnip greens (*Brassica rapa*) and mustard greens (*Brassica juncea*), or onion (*Allium cepa*)] and ranged between 10% and 22%. In contrast, CV for crops in which plant populations are small because of labor involved (tomato or cucumber) or space required [watermelon (*Cucurbita pepo*) or pumpkin (*Cucurbita maxima*)] were as high as 70%. Mean CV for tomato, cucumber, watermelon and pumpkin trials were 39%, 36%, 35%, and 48%, respectively (Table 5).

In general, CV values are even higher for weights within each grade because grade weights are fractions of total marketable yields. Yet, when market prices are much higher at the beginning of the season, growers may make higher profits with a small percentage of their total production. These results suggest two points. First, because of the inherent variability in plant yields, plot sizes commonly used may not allow CV values found in vegetable variety trials to be kept at or below the accepted levels for variety trials of particular commodities. The probability of detecting real mean differences among varieties can be low in experiments where responses have a high CV. Second, as commodities are grown for profit, early yields should be compared using market values.

QUESTION 8. Could yield data be published alone? Any attribute useful to distinguish or compare varieties may be measured. Traits measured might include yield, grade distribution, and horticultural attributes (Maynard, 2001). Some data commonly collected are crop-specific (Table 4). Yet, photosynthetic response (Bhagsari, 1990), plant nutritional characteristics (Quintana et al., 1996; Southwick et al., 1999), vitamin content (Simonne et al., 1997; Wang and Goldman, 1996), chemical composition (Kalt and McDonald, 1996), cooking tests (Paull et al., 2000), response to fertilizer rate (Mullins et al., 1999), consumer acceptance (Frank et al., 2001), taste tests (Brittain and McDonald, 1987; Simonne et al., 1999), disease reaction (Schultheis and Walters, 1998; Southwick et al., 1999), or post harvest behavior (Liang and Harbaugh, 2001) also provide useful information in assessing variety performance. It is, therefore, unlikely that yield alone would be sufficient to make a recommendation. Hence, yield alone should not be the sole data reported in manuscripts submitted to the Variety

Continued

Data acquisition	Observations
Yield expressed as 42-lb bushels for summer squash and cwt for winter squash and pumpkin, early yield for summer squash	
Days to first harvest	Plant habit
Number of fruit	
Fruit weight for pumpkin and winter squash	
Fruit shape	
Fruit color	
Early and total yield	Susceptibility to disease
Fruit weight	Occurrence of fruit defects
Fruit color	
Fruit shape	
Fruit firmness	
Percent soluble solids	
Yield	Susceptibility to disease
Days to first harvest	Occurrence of fruit defects
Fruit weight	
Proportion of extra-large, large, medium and small fruit	
Fruit shape	
Early and total yield	Susceptibility to disease
Days to first harvest	Rind characteristics
Average fruit weight	Internal characteristics
Percent soluble solids	

Trials category of *HortTechnology*.

When multiple attributes are measured on each plant, they constitute a multivariate response and there are advantages to consideration of techniques for multivariate analysis of variance (MANOVA) in testing hypotheses involving variety differences. To carry out analysis of variance (ANOVA) independently on a each of a sequence of attributes would ignore any information conveyed by associations among them. A MANOVA approach makes use of the possible, even probable, linear associations between attributes and may improve the efficiency of the analysis. An accessible review of MANOVA appropriate for use in variety trials can be found in Johnson and Wichern (1998; chapter 6). The authors include presentation of code and explanation of output using SAS. Costs associated with the MANOVA approach include a stronger reliance on normally distributed data and a more complex model and interpretation and discussion of the analysis. Methodology for multiple comparisons among multivariate variety means has not been developed thoroughly, though Bonferroni corrections are valid.

Question 9. When should data be corrected for stand? When assessing yield, stand count provides information on yield potential. A variety with poor establishment rate is unlikely to be acceptable. Therefore, it is not necessary to adjust for stand in most cases. Analysis of the stand rate itself may be of interest, but inflating yield by correction for stand would create an inaccurate measurement (Sullivan and Bliss, 1981). For example, bell pepper marketable yields were greater with stand reductions of 10%, 20%, or 30% compared to complete stands (Bracy, 1997). In a similar study with tomato, replanting improved marketable tomato yields only when the level of stand deficiency reached 30%, as compared to 10% and 20% stand deficiencies (Stoffella and Maynard, 1988). In some cases, however, when plant numbers are relatively large (such as in sweet corn variety trials) and plant stands are almost perfect (near 100%), it is possible to analyze yield data using covariance analysis. Stand is the covariate variable, and variety and blocks are the class variables.

Question 10. What units should be used to report data? The unit policy of *HortTechnology* detailed in the author instructions states that "authors submitting papers to be published in *HortTechnology* should use U.S. units followed by their metric equivalents in parentheses. If the original measurements or observations were made in metric units, report metric units first followed by their U.S. equivalents in parentheses. Authors may not use metric units without reporting their equivalents in U.S. units." Another classical issue involving units is the selection of the area used to report data: per plot or per acre? ASHS requires data to be reported on a per surface area (hectare or acre) basis. Therefore, results must be multiplied by a corrective factor to adjust for plot size. The corrective factor used is seldom reported, but may be easily calculated in plot length and between-row spacing are provided. However, it may prove useful for some readers to be able to compare results of different studies on a plot basis or on a linear bed foot basis. This can only be done if the corrective factor is provided, possibly in the Materials and Methods sections or as a footnote in the tables or figures.

Question 11. How can global indices be used to establish overall comparisons? A global index is a variable that is calculated based on the sum of ranks of other variables. Data collected on variety trials are usually

Table 5. Coefficient of variation (CV) observed in variety trials.[z]

Crop	Plot size[y]	Plants/ plot	No. of trials	Observed CV for marketable yield	
				Mean	Range
Broccoli (*Brassica oleracea* var *italica*)	20 ft (1st crop)	40	8	20	16–32
	20 ft (double crop)	40	3	37	19–66
Cabbage (*Brassica oleracea* var *capitata*)	20 ft (1st crop)	40	6	38	16–56
	20 ft (double crop)	40	1	53	---
Carrots (*Daucus carota*)	335 ft	6,700	1	10	---
Cucumber (*Cucumis sativus*)	20 ft	30	4	36	27–48
Garlic (*Allium sativum*)	2 rows × 5 ft	40	3	37	32–49
Green bean (*Phaseolus vulgaris*)	20 ft	20	2	48	30–65
Turnip (*Brassica rapa*) and mustard (*B. juncea*) greens	4 × 20 ft	1,200	4	16	12–22
Muskmelon (*Cucumis melo*)	30 ft	10	17	35	7–67
Okra (*Abelmoschus esculentus*)	10 ft	20	2	46	20–72
	20 ft	20	3	48	40–61
Onion (*Allium cepa*)	20 ft	40	1	21	---
	1 row × 50 ft	120	2	21	21
	4 rows × 40 ft	480	1	15	---
	4 rows × 20 ft	240	1	55	---
Ornamental corn (*Zea mays*)	4 rows × 20 ft	96	4	33	18–45
Bell pepper (*Capsicum annuum*)	20 ft	40	12	37	25–51
Hot pepper (*Capsicum annuum*)	5 ft	10	8	38	25–74
Pumpkin (*Cucurbita pepo, C. maxima, C. moschata*)	50 ft	20	25	48	9–90
Southern pea (*Vigna unguiculata*)	2 rows × 20 ft	40	2	45	44–45
Strawberry (*Fragaria ×ananassa*)	15 ft	30	3	14	10–20
Summer squash (*Cucurbita pepo*)	20 ft	13	12	29	8–63
Sweet corn (*Zea mays*)	6 rows × 23 ft	78	3	18	11–23
	2 rows × 20f t	48	16	30	7–66
Sweet potato (*Ipomoea batatas*)	30 ft	30	8	37	21–95
Tomato (*Lycopersicum esculentum*)	12 ft	8	12	39	22–70
Watermelon (*Citrullus lanatus*)	50 ft	10	11	37	14–61
	60 ft	12	7	29	15–45
Winter squash (*Cucurbita moschata*)	20 ft	20	2	29	18–39

[z] From trials in Simonne, 1996a, 1996b, 1997a, 1997b, 1998a, 1998b, 1999a, 1999b, 2000; Kemble 2000, 2001.

[y] 1 ft = 0.3 m.

analyzed using univariate statistics (analysis of variance, means comparison tests, non parametric tests). Yet, the recommendation of a variety is based on a global judgement that includes several attributes. For example, ear characteristics, and eating quality all contribute to the quality of sweet corn (Simonne et al., 1999). Stem length, diameter and vase life were used together to identify the overall best trachelium (*Trachelium caeruleum*) variety (Liang and Harbaugh, 2001). The evaluation of lemon (*Citrus limon*) varieties included fruit yield, juice yield and chemical composition, and tree survival rate (Fallahi et al., 1990). In many published articles, the overall evaluation is part of the discussion, and not of the statistical analysis. Evaluation is interpretative in nature, based on a set of individual measurements (Fallahi et al., 1990). In some articles where ranking procedures have been used, overall performance was based on a rank sum index (Liang and Harbaugh, 2001; Simonne et al., 1999). Global indices based on rank sums can be used to calculate the partial contribution of each attribute to the overall evaluation of that variety, thereby permitting the analysis of the contribution of each component trait.

Global indices are simple mathematical tools to define. Yet, some basic rules must be followed in establishing them. First, each entry has to be ranked for each attribute. Often, "1" is assigned to the entry with the highest mean, and "N" is assigned to the variety with the lowest (for a trial of N entries). It is essential that all rankings are oriented the same way in regard to desirability. Some variables such as yield represent desirable attributes. In this case, the higher the value, the better the variety. Other variables describe adverse or undesirable traits. This may be, for example, levels of bitterness in lettuce (*Lactuca sativa*), or cull weight. In this case, the higher the value, the less desirable the attribute. Then, when results from trials from different years/location are used to define a global index, it is not uncommon for the number of ranks to be different (because of a different number of entries at each year/location). In this case, attention has to be paid to the way the ranks are assigned so that no bias is introduced. Ranks should be assigned in a way that keeps the rank sum of all tests and attributes the same. Otherwise, more weight is given to the attribute with higher rank sum, thereby introducing some bias. Finally, when two or more means are numerically the same, then a tie in rank occurs. The proper way to handle two- way ties at rank p, is to assign twice the rank p + 1/2. The following rank is then p + 2. When a three-way tie occurs at rank p, the rank p+1 should be assigned three times. The following rank would then be p+3. This procedure allows the sum of the ranks to be constant, despite the presence of ties. The three sums p + (p+1) + (p + 2), (p + 1/2) + (p + 1/2) + (p + 2), and (p + 1) + (p + 1) + (p + 1) are equal to (3p + 3). Failure to properly handle ties usually ends up modifying the sum of the ranks, and introduce some bias. Inspection and description of these indices can be informative, but how to use them for statistical inference about variety differences is not clear. Because of their dependence structure, ranks or rank sums and hence global indices have different sampling properties than do statistics obtained from independent observations. Any single attribute within a year–location can be analyzed using classical nonparametric methods based on ranks such as Kruskal-Wallis (one-way) or Friedman (block designs) described in Hsu (1996) or Hollander and Wolfe (1999). Inference based on global indices taken as the sum of dependent ranks from incomplete blocks is a problem that requires further study.

An alternative to the rank-sum approach for analyzing this highly dimensional response through a single index, or possibly a small number of indices, is principal component analysis (PCA). In PCA, the first principal component is the linear combination of the attributes that explains the highest proportion of of the sum of the sample variances of all the components, subject to a constraint on how big the coefficients in the linear combination can be. The second principal component is found similarly, with the constraint that it be uncorrelated with the first, and so on. If a sample is transformed via the first principal component to a single dimension, then an interpretable ordering of the varieties may appear, such as a yield index, or an index of consumer preference of eating quality. It is conceivable that a univariate analysis of variance on these dimension-reduced transformations might also be enlightening. PCA is popular in many areas of agricultural research and may be appropriate for study in variety tests as well.

Issues related to statistics and experiment design

Question 12. How much detail is required to describe the statistical methodology used? Guidelines for improved presentation of ANOVA and regression results are provided by Wehner and Shaw (1994) and should be adopted for the scholarly publication of variety trials in *HortTechnology*. In particular, these authors applaud the inclusion of ANOVA tables in manuscripts. This should be standard practice. Without a complete ANOVA table, it is difficult to understand all of the sources of variability (such as block, treatment, or time) in an experiment, and to what degree they explain variation in the response variables. In particular, the mean square for error (MSE) term should always be reported, as it estimates error for replications within a block × variety combination. The MSE term is central to all subsequent inference, including tests and confidence intervals. MSE can also provide a reference point for precision in comparison of multiple studies. As done routinely in other refereed publications, the experiment design should also be decribed in the materials and methods section of manuscripts submitted to the Variety Trials category of *HortTechnology*.

In block designs, standard errors for differences of variety means are the same whether block effects are treated as fixed or random. However, in random block models, variety means by themselves involve averages of correlated effects and so have standard errors that are different than in the fixed block effect models. Proc Mixed in SAS can easily be used to obtain the appropriate standard errors:

```
Proc Mixed;
    Class variety block;
    Model y = variety;
    Random block;
    Lsmeans variety/stderr;
Run;
```

Question 13. Should evidence that the assumptions for a legitimate use of ANOVA be included in a manuscript? Some mention of whether or not the data conform, at least approximately to the statistical assumptions underlying ANOVA techniques ought to be included to validate the statistical methodology. ANOVA techniques assume normally distributed data with homogeneity of variance. While it is has been established that the F test enjoys robustness to a variety of departures from normality, this does not imply that it is valid for any distribution. Similarly, mean separation procedures are based upon assumptions of independent and normally distributed data. Most statistics texts advocate inspection of residual plots or goodness-of-fit statistics. Inclusion of these plots may not be appropriate for manuscripts here, but some mention that they were inspected would be reassuring to readers. It is straightforward to generate residuals from any model and use them in diagnostic plots. For example, the output statement in the SAS code below creates a temporary SAS data set containing the original data along with the fitted values $\hat{y}$ for a response variable y and the residuals $e = y - \hat{y}$. These residuals can then be plotted against the fitted values to check for homogeneity of variance (Bartlett's test may also be used for this purpose). Normal plots for the residuals or goodness-of-fit statistics can then be used to assess the assumption of normality.

To the greatest extent possible, data and analyses should be clarified for ease of understanding and for possible use in future work. There may be considerable overlap among multiple experiments or publications and all-important statistical power for finding variety differences could be gained by pooling results. Ensuring high standards in presentation and publication will do much towards this end. Even better (although seldom done in past agricultural scientific literature) would be inclusion of the steps followed when software is used. An example of inclusion of SAS code would be the following:

```
Proc Glm;
    Class variety block;
    Model yield = variety block;
    Lsmeans variety/cl pdiff = all adjust = tukey;
    Output out = resdata r = residuals p = fitted;
Run;

Proc Plot;
    Plot residuals*fitted;
Run;

Proc Univariate normal plot;
    Var residuals;
Run;
```

Another possibility is to include a URL pointing readers to locations on the internet where data and SAS or other software code can be found.

Question 14. Reporting raw data, ranks, or percentage of check within replication: what are the pros and cons? If data do not approximately conform to Gaussian (normal) distributional assumptions underlying ANOVA then logarithmic, exponential or power transformations may alleviate the problem. When data transformation fails, ranks of responses can be used in nonparametric analyses for comparisons of variety medians. Nonparametric methods are generally less efficient than ANOVA for normally distributed data, so that ranks should only be used when necessary. Drawbacks to subtracting means or using ratios of measurements to some standard include a loss of degrees of freedom for estimation of experimental error variance. In some cases, the difference from some reference measurement point really is an appropriate response to analyze.

Question 15. Are single year–location trials as well as nonreplicated data publishable? Inference is the legitimate claim that the results observed on a sample also apply to the population from which that sample was drawn. In order to control its level of statistical risk, inference is based on an assessment of experimental repeatability in time and space. Multiple locations allow the estimation of the variety × environment (location) interaction. In most trials, this interaction is significant, thereby indicating that the performance of varieties differs from location to location (Hodges et al., 1995; Poysa et al., 1986). When nonreplicated data are collected, a broad inference cannot be made since no estimate of variance is available. (As discussed above, making this inference is irrelevant when the goal of the trial is to eliminate the worst varieties due to an unacceptable intensive attribute). If block effects (such as fertility or water gradients in soil) are strong and can conceivably affect different varieties differently, then replication within blocks is needed to estimate these interactions. If it were reasonable to believe that all varieties benefit to the same degree from block effects, then nonreplicated data would be acceptable. If interactions exist, then the nonreplicated randomized block design (RBD) model is underspecified, and estimates for variety differences can be biased. It should be noted that replicated data may be collected from single-plot trials only when intensive variables are measured. Intensive variables such as growth habit, disease resistance, fruits type and shape, may be collected more than once on a single plot. Except in this situation, it is unlikely that nonreplicated data will be acceptable in manuscripts submitted to the Variety Trials category of *HortTechology*. In some limited cases, single-location trials may be acceptable, especially when the environmental conditions are relatively controlled such as in greenhouse studies.

The same issues arise when considering whether or not replication across times is required for statistical inference about variety differences. If assumptions about seasonal stability or additivity are plausible, then inference from single year trials is reasonable. However, variety × weather or variety × year interactions would seem to be of tremendous interest for many varieties. Suppose for example that a variety is clearly identified as having higher mean yield than others in a single-year trial. It is not possible to infer that it will also be highest in another year with different growing conditions. For inference about variety × weather or variety × year interactions, single-year trials are not sufficient.

Question 16. What are the most appropriate mean separation techniques for mean comparisons in variety trials? A survey of the voluminous literature on the topic indicates that the multiple comparisons issue is a controversial one. The collection of papers defies enumeration (Chew, 1976; Gates, 1991; Little, 1978; Mihail and Black, 1991; Saville, 1990; Swallow, 1984; Tukey, 1991). There have been a number of simulation studies of multiple comparison procedures (MCPs) or mean separation procedures (Carmer and Swanson, 1973; Einot and Gabriel, 1975). Finally, most textbooks on statistical methods address the issue and some (Hochberg and Tamhane, 1987; Hsu, 1996) are devoted entirely to it. This problem will be reviewed here and insight into pertinence in variety trials will be discussed.

Inference and types of error. As mentioned at the outset, the quality of a publication depends upon the scope and strength of the inference that can be made from the analysis of the experiment. Occasionally, erroneous conclusions will be drawn from analyses in the form of perceived differences among equivalent varieties (type I), real differences among varieties that go undetected (type II), or to declare the wrong ordering of real differences of variety (type III). This last error would occur when the sample mean yield for variety A is sufficiently larger than variety B to declare significance, but variety B actually has a higher long-run (population) mean yield.

The statistical problem is to control or minimize these error rates. MCPs have been developed for experiments that attempt to answer many questions at once; they account for multiplicity of comparisons. Hsu (1996) classifies MCPs into five possible categories of strength of inference. The weakest of these categories is individual comparison methods. These MCPs make no adjustment for multiplicity and do not guarantee the simultaneous correctness of multiple assertions of variety differences with any confidence level. The strongest is simultaneous confidence interval methods. These MCPs assert ranges of plausible values for variety differences, and guarantee them to have simultaneous correctness with specified confidence levels.

Control of types of errors by MCPs. Erroneous assertions are unavoidable, but MCPs have been developed to control the type I error rate while accounting for multiplicity. When making many comparisons among varieties, the experimentwise error rate is defined as the average proportion of experiments in which at least one type I error is committed. The comparisonwise error rate is the average proportion of comparisons in which a type I error is committed. Some MCPs, such as Tukey's procedure, sometimes called the honestly significant difference, or the Tukey-Welsch procedure (denoted REGWQ in SAS) are constructed to control for the experimentwise error rate. Other MCPs, such as the least significant difference (LSD), the protected LSD, or Duncan's multiple range test are not.

MCP for Hsu's weakest level of inference. Persuasive arguments are made in Saville (1990) that only individual comparisons are needed in experiments such as variety trials, that stronger forms of inference are too complex for interpretation, can lead to inconsistencies in declaring differences significant and suffer from a high type II error rate. So, Hsu's weakest form of inference can be achieved simply by comparing any observed difference to a least significant difference or LSD based on the t distribution of any estimated, standardized difference of sample means. If it is acceptable to accept a comparisonwise error rate of $\alpha = 0.05$, without undue concern for experimentwise error rates, then the LSD procedure is a reasonable recommendation.

A scan of current articles indicates

Table 6. Number of possible pairwise comparisons and average number of errors committed with an increasing number of varieties in a trial.

No. of entries	No. of comparisons	Avg no. of errors committed[z]
5	10	0.5
10	45	2.3
20	190	9.5
30	435	21.8
40	780	39

[z]Average number of confidence intervals that will miss the true differences.

that this is a popular technique in the horticultural literature, easy to explain and report in tables of means. Confidence intervals can also be constructed without adjustment for multiplicity. Their interpretation is that 95% of them will cover the true variety differences. If this is an acceptable level of confidence, then no adjustment for multiplicity is warranted, so long as the limitation of the strength of the inference is mentioned in the analysis. For example, Table 6 provides the number of pairwise comparisons needed in a variety trial with between 5 and 40 entries and the average number of confidence intervals that will miss the true differences. One well-known study (Carmer and Swanson, 1973) reported the following Monte Carlo estimates of experimentwise error rates for randomized block designs with k = 5, 10 or 20 number of treatments (entries) and varying numbers of replications n = 3, 4, 6, 8. The Monte Carlo standard errors for the estimated experimentwise error rates appear in parentheses (Table 7) and are determined from the fact that 4000 simulations were used for each treatment configuration. This indicates that in trials with k = 20 equal treatments if Duncan's procedure is used with comparisonwise error rate $\alpha = 0.05$, there will be false discoveries of differences among varieties in about 61% to 64% of these types of experiments (Table 7). The example is rather extreme, as few would attach much credence to an omnibus equality of all varieties in many variety trials in the first place, but the same thing happens for many configurations in which there are some equalities among variety means.

MCP for Hsu's strongest level of inference. If stronger inference is desired, then a simple and highly informative analysis of variety trials that addresses the multiplicity issue is to obtain simultaneous confidence intervals for all pairwise differences with the property that the chance that they all cover the true mean differences is 95%. This method is easy to implement using SAS or other statistical or spreadsheet software. The sample code provided earlier will report the intervals for all pairwise comparisons. Space considerations would make it difficult to report all of these in an article, particularly when there are many attributes in addition to yield under consideration. Only those intervals that are of interest to growers or researchers need to be reported. Consider an example taken from a variety trial in Michigan to evaluate six varieties of early season strawberry. Assume the actual means [in 100-lb/acre increments (112.1 kg·ha^{-1})] that would be achieved with infinite sample size are $\mu_1 = 40$, $\mu_2 = 40$, $\mu_3 = 50$, $\mu_4 = 90$, $\mu_5 = 90$, and $\mu_6 = 110$. Assume further that a complete randomized block design is used, with a total of 36 observations. The block effects were assumed to be modest: with two blocks yielding on average 2000 lb/acre (2,240 kg·ha^{-1}) more, two blocks yielding 2000 lb/acre less, and two blocks yielding the average. Data simulated under this model assuming that the standard deviation for replications of each variety is about $\sigma = 20$ appear in Table 8. The SAS code used to generate these data were

```
Data One;
  Array alpha {6} (–30, –30, –20, 20, 20, 40);
  Array beta {6} (–20, –20, 0, 0, 20,  20);
  Do block = 1 to 6;
    Do variety = 1 to 6;
      y = 70 + alpha {i} + beta {j} + 20 × normal (2);
      y = round (y, 2);
      Output;
    End;
  End;
Run;
```

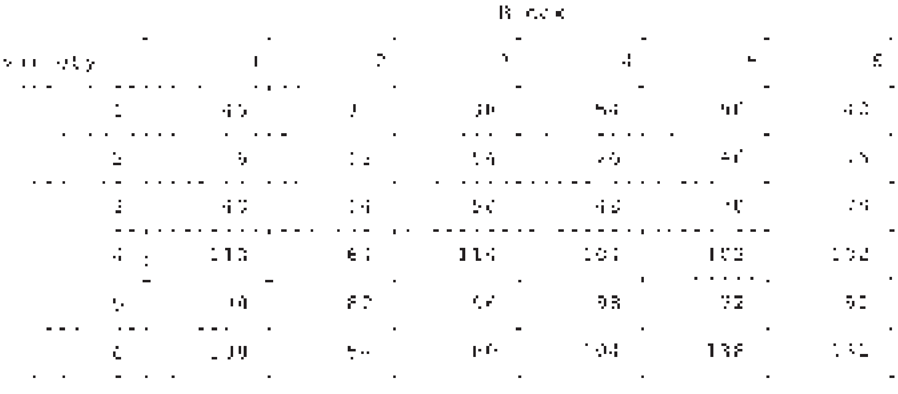
The output with the ANOVA table and all differences was produced by SAS and cut and directly pasted into this document (Table 9). The only differences in varieties illuminated by the analysis are those involving pairs (1, 4), (1, 5), (1, 6), (2, 4), (2, 5), (2, 6), (3, 4), (3, 5), and (3, 6), though the (1, 3) and (1, 5) differences contain plausible differences as small as 700 or 800 lb/acre (785 or 897 kg·ha^{-1}). Since the true variety means are assumed to be known, the error rate from this particular simulated experiment is known. No type I errors occurred. The differences involving pairs (1, 3), (2, 3), (4, 6), (5, 6) were all missed, so that four type II errors resulted. Note that the magnitude of these 4 errors is very small relative to other differences among the varieties, so that there may be smaller cost associated with missing them. Listing the simultaneous confidence intervals provides more information about the precision with which the effects can be estimated. An interval estimate for the difference between varieties 2 and 6 is 4,400 and 10,800 lb/acre (4,932 and 12,105 kg·ha^{-1}). Reporting the interesting confidence intervals may be more informative than tables with the requisite "means followed by the same letter do not differ significantly," a message that Little (1978) belittles. These same data could also be analyzed without regard to multiplicity [6 (6 – 1)/2 = 15 comparisons] using least significant difference (LSD). The output was again produced by SAS and cut and pasted into this document (Table 10). Note the warning at the bottom of the SAS output. Note also the type I error that occurred when comparing varieties 1 and 2. The observed comparison wise error rate here is then 1/15 = 6% while the observed experimentwise error rate is 100%. Hayter and Hsu (1994) reported that Tukey's procedure applies to randomized block designs and to balanced designs, but for unbalanced incomplete block designs, it is currently only conjecture that the Tukey MCP preserves the experimentwise error rate for more than k = 3 varieties. In this last

Table 7. Observed experimentwise type I error rates for three common multiple comparison procedures for increasing number of entries.

Multiple comparison procedure	No. of entries (k)		
	5	10	20
Tukey	5%	4.8%	4.7% (0.003)[z]
Duncan	18.2%	37.3%	62.6% (0.008)
LSD	25.6%	58.4%	89.5% (0.008)

[z]Monte Carlo standard error for the estimated experimentwise error rates (adapted from Carmer and Swanson, 1973).

Table 8. Simulated data on strawberry early yield [in 100-lb/acre (112.1 kg·ha^{-1}) increments] for a randomized block design used as an example.[z]



	Block					
Variety	1	2	3	4	5	6
1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4	113	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
6	[illegible]	[illegible]	[illegible]	[illegible]	138	[illegible]

[z] Six varieties and six blocks were used in this example.

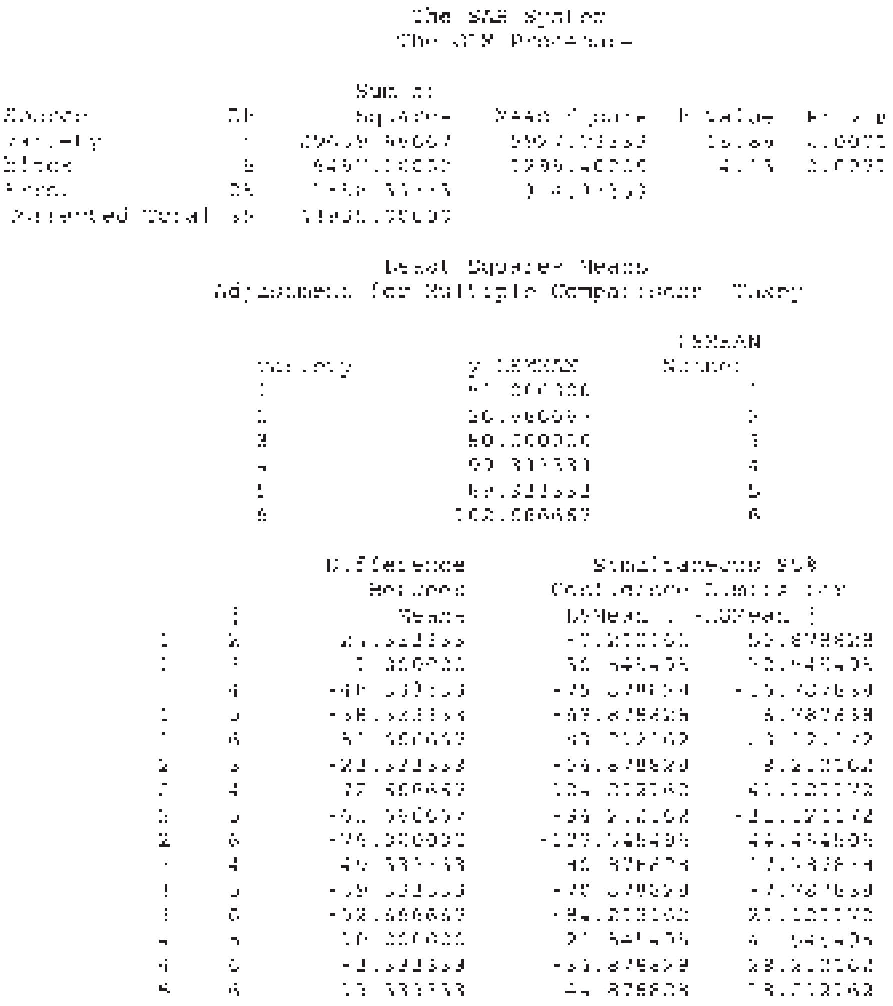
case, a Bonferroni adjustment or simulation approach is recommended. Both are easily provided for in SAS using **adjust = bon** or **adjust = simulate** in the LSMEANS statement. Theoretical details can be found in Westfall et al. (1999).

Trade-off in error types. Controlling experimentwise error rate comes at a price. There is generally a type I for type II error trade-off. This alone is perhaps a strong argument not to adjust for multiplicity. Without attaching some loss or cost to the two types of errors or incorporating production costs associated with actual varieties that could be used in conjunction with interval estimates to attempt to make decisions based on profitability, there is no immediate reason why controlling for type I error rate is more important than minimizing type II error rate. This is perhaps unusual in that many areas where multiple comparisons might be used, in medical or pharmaceutical applications for example, type I errors are often more grievous than type II errors. The response to the next question in the list attempts to address power considerations. For now, some power can be gained while controlling for experimentwise error rate when it is not necessary to make all pairwise comparisons.

MCP and variety testing. Sometimes inference can be restricted to questions concerning a control or identification of a best variety and a search for inferior varieties relative to this unknown best. MCPs that are useful in variety trials can be classified into three groups: 1) all-pairwise comparisons (MCA); 2) multiple comparisons with a control (MCC), and 3) multiple comparisons with the best (MCB). MCA was discussed previously. Dunnett's procedure can be used for MCC problems. It preserves the experimentwise error, is more powerful than MCA procedures and is easy to implement using SAS. In the strawberry example, suppose that interest lies in comparisons involving variety 4 so that it is a standard or reference variety or control. The following LSMEANS statement within Proc Glm will get SAS to carry out Dunnett's MCP: **lsmeans variety/pdiff = control ('4') adjust = dunnett**

Inspection of the output (Table 11) reveals that only one nonzero difference (4, 6) was missed, and precision is greater. These intervals are more narrow than the ones necessary for the MCA problem. From this analysis, the standard appears to produce significantly more than varieties 1, 2, and 3 and substantially more than varieties 1

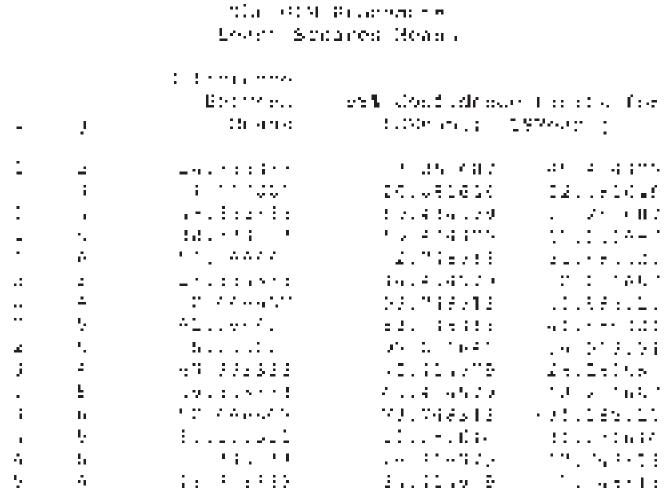
Table 9. SAS output for analysis of variance (ANOVA) for simulated strawberry data with population variety mean yields of 4000, 4000, 5000, 9000, 9000, and 11000 lb/acre (1lb/acre = 1.12 kg·ha^{-1}) for varieties 1 to 6, respectively, in a randomized complete block design (RCBD).[z]



```
                          The SAS System
                         The GLM Procedure

                          Sum of
Source            DF      Squares     Mean Square   F Value   Pr > F
variety            5   [illegible]    [illegible]  [illegible] [illegible]
block              5   [illegible]    [illegible]  [illegible] [illegible]
Error             25   [illegible]    [illegible]
Corrected Total   35   [illegible]

                       Least Squares Means
            Adjustment for Multiple Comparisons: Tukey

                                          LSMEAN
           variety        y LSMEAN        Number
           1              [illegible]          1
           2              26.666667            2
           3              50.000000            3
           4              [illegible]          4
           5              [illegible]          5
           6              102.666667           6

                      Difference          Simultaneous 95%
                       Between          Confidence Limits for
          i    j        Means           LSMean(i) - LSMean(j)
     1    2      [illegible]       [illegible]    [illegible]
     1    3      [illegible]       [illegible]    [illegible]
     1    4      [illegible]       [illegible]    [illegible]
     1    5      [illegible]       [illegible]    [illegible]
     1    6      [illegible]       [illegible]    [illegible]
     2    3      -23.333333        [illegible]    [illegible]
     2    4      [illegible]       [illegible]    [illegible]
     2    5      [illegible]       [illegible]    [illegible]
     2    6      -76.000000        [illegible]    [illegible]
     3    4      [illegible]       [illegible]    [illegible]
     3    5      [illegible]       [illegible]    [illegible]
     3    6      -52.666667        [illegible]    [illegible]
     4    5      [illegible]       [illegible]    [illegible]
     4    6      [illegible]       [illegible]    [illegible]
     5    6      [illegible]       [illegible]    [illegible]
```

[z] See Table 8 for corresponding data set. Simultaneous inference for pairwise comparisons among sample variety means uses Tukey's procedure. No type I errors committed.

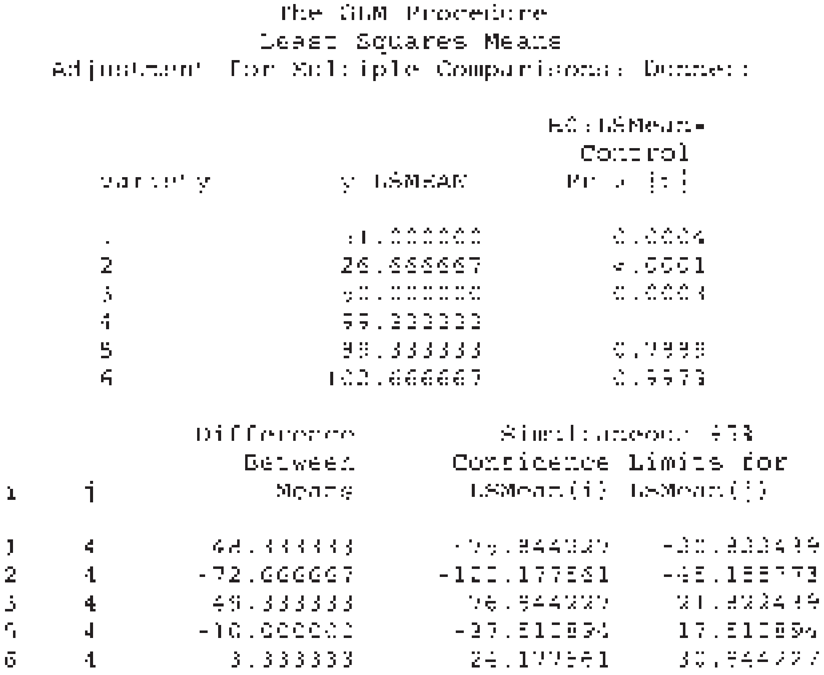
Table 10. SAS output for multiple pairwise comparisons among sample variety means from simulated strawberry data with population variety means of 4000, 4000, 5000, 9000, 9000, and 11000 lb/acre (1lb/acre = 1.12 kg·ha^{-1}) for varieties 1 to 6, respectively, in a randomized complete block design (RCBD).[z] To ensure overall protection level, only probabilities associated with preplanned comparisons should be used.



```
[illegible]
```

[z]See Table 8 for corresponding data set. No statistical adjustment is made for multiplicity of comparisons. To declare the difference between means 1 and 2 as statistically significant is to commit a type I error.

Table 11. SAS output for multiple comparisons with a control (variety 4) using Dunnett's procedure for simulated strawberry data in a randomized complete block design (RCBD).[z]



```
                        The GLM Procedure
                       Least Squares Means
         Adjustment for Multiple Comparisons: Dunnett

                                                  H0:LSMean=
                                                    Control
           variety           y LSMEAN             Pr > |t|

           1                51.000000               0.0004
           2                26.666667               <.0001
           3                50.000000               0.0004
           4                99.333333
           5                89.333333               0.7998
           6               102.666667               0.9973

                     Difference               Simultaneous 95%
                        Between            Confidence Limits for
i      j                  Means            LSMean(i)-LSMean(j)

1      4             -48.333333          -75.844227   -20.822439
2      4             -72.666667         -100.177561   -45.155773
3      4             -49.333333          -76.844227   -21.822439
5      4             -10.000000          -37.510894    17.510894
6      4               3.333333          -24.177561    30.844227
```

[z]See Table 8 for corresponding data set. True mean yields for varieties 1 to 6 were 4000, 4000, 5000, 9000, 9000, and 11000 lb/acre (1 lb/acre = 1.12 kg·ha^{-1}), respectively. True differences of control versus varieties 1, 2, and 3 were detected.

and 2 while variety 6 may produce as high as 5,200 lb/acre (5,828 kg·ha^{-1}) more than the standard. Lastly, inference may be only for identification of a best variety with respect to a single attribute at a time, such as yield. Hsu (1996) has developed an MCB procedure that has not been enabled yet in Proc Glm, but is available in SAS/ INSIGHT. This procedure is not as commonly used yet, but seems appropriate for variety trials.

QUESTION 17. Is there a method to determine the adequate number of replications needed? Should power be reported? When designing the experiment, both types of error rates should be considered: false discovery or accidental declaration of equivalent varieties to be different (type I) and failure to declare sample differences among substantially different varieties to be significant (type II). Researchers have rightly criticized undue emphasis on type I error rates. Indeed, *P* values and hypothesis tests are formulated to control the probability of false discovery, but many researchers point out that false discovery is no more severe an error as failure to discover. Power has been insufficiently addressed in many fields, perhaps because it is inconvenient to calculate. Software that will compute power is much less prevalent and less well-known and few practitioners are familiar with appropriate packages. Power considerations also require the practitioner to be more specific and to elaborate about the questions being investigated. As is often the case when a client asks a question to a statistician, the client gets several questions back in return. How much of a difference between varieties is meaningful or is it possible to hypothesize a meaningful configuration of variety means a priori? What procedure will be used to address the problem of multiplicity of comparisons? How high must the probability be to detect a given variety difference or what proportion of actual differences declared significant is acceptable?

Sample size computations are straightforward for the F test in one-way ANOVA (see Rao, 1998, chapter 9.8 for example). Such a computation requires only the specification of an effect size and a guess at the experimental standard deviation of a response. This approach provides a start for a variety trial. As an example, suppose that six varieties of early season strawberry are to be evaluated in a variety trial and the actual means in 100-lb/acre increments that would be achieved with infinite sample size for the six varieties are about $\mu_1 = 40$, $\mu_2 = 40$, $\mu_3 = 50$, $\mu_4 = 90$, $\mu_5 = 90$, and $\mu_6 = 110$. The effects, or differences from average, of the six varieties are then $\alpha_1 = 30$, $\alpha_2 = -30$, $\alpha_3 = -20$, $\alpha_4 = 20$, $\alpha_5 = 20$, and $\alpha_6 = 40$. Suppose that the standard deviation for replications of each variety is $\sigma = 20$. Suppose $n = 3$ replications will be made in a completely randomized design. Following Rao (1998), the F ratio will have a noncentral F distribution with $v_1 = 5$ and $v_2 = 12$ as numerator and denominator degrees of freedom, respectively, and a noncentrality parameter of $2\lambda = n\sum a_i^2/\sigma^2 = 34.5$. The area to the right of the $\alpha = 0.05$ critical value F (0.95, 5,12) = 3.1059 under this noncentral F distribution is 0.9680. The SAS code below computes this power:

```
Data One;
    Lambda = 3* (30**2 + 30**2 + 20**2 + 20**2
      + 20**2 + 40**2)/400;
    Fstar = finv (0.95, 5, 12);
    Power = 1-probf (fstar, 5, 12, lambda);
Run;
```

Beyond this simple computation, another power consideration involves the proportion of the real differences that are expected to be detected. In the example above, of the 15 pairwise variety differences, 13 are nonzero. If a MCA procedure is used, what proportion of these 13 nonzero differences is expected to be found? One approach to answering this question is simulation. A Monte Carlo estimate of the proportional power can be constructed from averaging the proportion of 13 differences detected over many simulations. The more simulated data sets, the more accurate the estimate of power. The SAS macro %SimPower developed by Westfall et al. (1999) accomplishes this easily and allows for each of MCA, MCC, and MCB types of comparisons using the Tukey, Dunnett, or Tukey-Welsch procedures. The macro is currently available at <http://ftp.sas.com/samples/A56648> (Westfall et al., 1999). Once the macro has been compiled, the following statement is all that is needed: %SimPower (method = tukey, n = 4, s = 20, truemeans = (40, 40, 50, 90, 90, 110), nrep = 100, seed = 123);

The output from this invocation of the macro appears below:

```
Method = TUKEY, Nominal FWE = 0.05, nrep
= 100, Seed = 123
  True means = (40, 40, 50, 90, 90, 110), n = 4,
  s = 20
Quantity            Estimate     95% CI
Complete Power      0.00000   (0.000, 0.000)
Minimal Power       0.98000   (0.953, 1.000)
Proportional Power  0.49308   (0.460, 0.526)
```

So that with this design and this a priori specification of variety effects and error, about half of the real differences will be detected, while the chance of rejecting the overall hypothesis of no variety effects is estimated as 0.98% (The widths of the confidence intervals can be decreased by increasing nrep when the macro is called). In publishing the findings from variety trials, computations like this would illuminate the possible limitations of the experiment to find all differences. The shortcoming of this software is lack of functionality beyond one-way models. Most variety trials are block designs and require two-factor ANOVA. In the absence of block × variety interaction, many designs, including balanced randomized block designs have what is called a one-way structure, and theory for MCPs carries over from one-way models to general linear models. Thus, the MCPs and power simulators apply to some more general models. For models that do not have a simple structure, such as unbalanced incomplete block designs, special provisions should be made to enable computation of power under meaningful alternatives.

Conclusion

The discussions presented here are intended to assist in establishing guidelines for the important issues related to the publication of varieties trial results. Without adequate scientific and scholarly standards, inferences drawn from variety trials can be argued to be invalid. However, as in all scientific work, the decisions on how to conduct variety trials are essentially the authors' prerogative. In this discussion, a limited number of issues was perceived as essential in order to maintain publication standards expected from ASHS publications: objectives stated clearly, plots of adequate size, reporting ANOVA table and MSE, inspecting the status of ANOVA assumptions, using nonparametric statistics when appropriate, and ending with a clear conclusion. Because of the large number of statistical questions asked from limited data, these are challenging statistical issues in the analysis of variety trials. Presentation of statistical power was also taken into consideration. Some decisions must be made regarding the importance of erroneous inference. If multiple type I errors are acceptable, then arguments can be made for using procedures such as LSD that make no adjustment for multiplicity. The benefit of this weaker form of inference is a gain in statistical power. However, it is argued here that tests alone may not be suffiencient or even useful for readers of these articles. Interval estimation of differences among varieties, mean yields for example, may better enable readers to select or recommend varieties after taking into account production costs. Simultaneous correctness with a specified level of confidence is also clearly desirable in these interval estimates. Hopefully, this discussion will stimulate authors and reviewers, as well as readers, and will ultimately result in manuscripts of high quality.

Literature cited

Bhagsari, A.S. 1990. Photosynthetic evaluation of sweetpotato germplasm. J. Amer. Soc. Hort. Sci. 115(4):634–639.

Bracy, R.P. 1997. Bell pepper yields not affected by stand deficiency or replanting. HortTechnology 7(2):138–142.

Brittain, M.J. and N.A. McDonald. 1987. Techniques used in performance trials of celery, leeks, parsnips and sweet corn. J. Natl. Inst. Agr. Bot. 17:345–352.

Brown, G.K., D.E. Marshall, B.R. Tennes, D.E. Booster, P. Chen, R.E. Garrett, M.O. O'Brien, H.E. Studer, R.A. Kepner, S.L. Hedden, C.E. Wood, D.H. Lenker, W.F. Miller, G.E. Rehkugler, D.L. Peterson, and L.N. Shaw. 1983. Status of harvest mechanization of horticultural crops. Amer Soc. Agr. Eng., St. Joseph, Mich.

Carmer, S.G. and M.R. Swanson. 1973. Evaluation of ten pairwise MCPs by Monte-Carlo methods. J. Amer. Stat. Assn. 68(341):66–74.

Chew, V. 1976. Comparing treatment means: a compendium. HortScience 11(4):348–356.

Conover, W.J. and R.L. Iman. 1981. Rank transformations as a bridge between parametric and nonparametric statistics. Amer. Stat. 35(3):124–129.

Cramer, C.S. 2001. Comparison of open-pollinated and hybrid onion varieties for New Mexico. HortTechnology 11(1):119–123.

Crossa, J. 1990. Statistical analysis of multilocation trials. Adv. Agron. 44:55–85.

Einot, I. and K.R. Gabriel. 1975. A study of the powers of several methods of multiple comparisons. J. Amer. Stat. Assn. 70(351):574–583.

Eskridge, K.M and R.F. Mumm. 1992. Choosing plant varieties based on the probability of outperforming a check. Theor. Appl. Genet. 84:494–500.

Fallahi, E., D.R. Rodney, and Z. Mousavi. 1990. Growth, yield, and quality of eight lemon varieties in Arizona. J. Amer. Soc. Hort. Sci. 115(1):6–8.

Fernandez, G.C.J. 1991. Analysis of genotype x environment interaction by stability estimates. HortScience 26(8):947–950.

Frank, C.A., R.G. Nelson, E.H. Simonne, B.K. Behe, and A.H. Simonne. 2001. Consumer preference for color, price, and vitamin C content of bell peppers. HortScience 36(4):795–800.

Gates, C.E. 1991. A user's guide to misanalyzing planned experiments. HortScience 26(10):1262–1265.

Hayter, A.J. and J.C. Hsu. 1994. On the relationship between stepwise decision procedures and confidence interval estimates. J. Amer. Stat. Assn. 89(425):128–136.

Hochberg, Y. and A.C. Tamhane. 1987. Multiple comparison procedures. Wiley, New York.

Hodges, L, D.C. Sanders, K.B. Perry, K.M. Eskridge, K.M. Batal, D.M. Granberry, W.J. McLaurin, D. Decoteau, J. Dufault, J.T. Garrett, and R. Nagata. 1995. Adaptability and reliability of four bell pepper varieties across three southeastern states. HortScience 30(6):1205–1210.

Hollander, M.. and D.A. Wolfe. 1999. Nonparametric statistical methods. 2nd ed. Wiley, New York.

Hsu, J.C. 1996. Multiple comparisons: Theory and methods. Chapman and Hall. London.

Johnson, R.A. and D.W. Wichern. 1998. Applied multivariate statistical analysis. 4th ed. Prentice-Hall, Upper Saddle River, N.J.

Kalt, W. and J.E. McDonald. 1996. Chemical composition of lowbush blueberry varieties. J. Amer. Soc. Hort. Sci. 121(1):142–146.

Kemble, J.M. (ed.). 2000. Spring 2000 commercial vegetable variety trials. Auburn Univ., Ala., Reg. Bul. 5.

Kemble, J.M. (ed.). 2001. Fall 2000 commercial vegetable variety trials. Auburn Univ., Ala., Reg. Bul. 6.

Kessler, J.R., Jr., J.L. Sibbley, B.K. Behe, D.M. Quinn, and J.S. Bannon. 2000. Herbaceous perennial trials in central Alabama. HortTechnology 10(1):222–228.

Liang, R. and B.K. Harbaugh. 2001. Evaluation of *Trachelium* varieties as cut flowers. HortTechnology 11(2)316–318.

Little, T.M. 1978. If Galileo published in HortScience. HortScience 13(5):504–506.

Maynard, D.N. 1987. Vegetable variety evaluation demonstrations: A manual for county extension faculty. Univ. Fla. IFAS Coop. Ext. Ser. Circ. 762.

Maynard, D.N. 2001.Variety selection, p. 15. In: D.N. Maynard and S.M. Olson (eds.). Vegetable production guide for Florida. Univ. Fla. IFAS, Gainsville.

Mihail, J.D. and T.L. Black. 1991. Comparison of treatment means: A statistical fantasy. J. Nematol. 23(4S):557–563.

Morales, M.R. and L. Maynard. 2000. Midwestern vegetable variety trial report for 2000. Purdue Univ., West Lafayette, Ind., Bul 798.

Mullins, C.A. and R.A. Straw. 2001. Performance of filet-type snap bean variety in Tennessee. HortTechnology 11(1):124–127.

Mullins, C.A., R.A. Straw, B. Pitt, Jr., D.O. Onks, M.D. Mulles, J. Reynolds, and M. Kirchner. 1999. Response of selected sweet corn varieties to nitrogen fertilization. HortTechnology 9(1):32–35.

Orzolek, M.D., W.J. Lamont, and L. Otjen. 2000. 1997 spring and fall cabbage variety trials in Pennsylvania. HortTechnology 10(1):218–221.

Paull, R.E., G. Uruu, and A. Arakaki. 2000. Variation in the cooked and chipping quality of taro. HortTechnology 10 (4):823–829.

Poysa, V.W., R. Garton, W.H. Courtney, J.G. Metcalf, and J. Muehmer. 1986. Genotype–environment interactions in processing tomatoes in Ontario. J. Amer. Soc. Hort. Sci. 111(2):293–297.

Quintana, J.M., H.C. Harrison, J. Nienhius, J.P. Palta, and M.A. Grusak. 1996. Variation in calcium concentration among sixty Si families and four varieties of snap bean (*Phaseolus vulgaris* L.). J. Amer. Soc. Hort. Sci. 121(5):789–793.

Rao, P.V. 1998. Statistical research methods in the life sciences. Brooks/Cole, Pacific Grove, Calif.

Rowell, B. (ed.). 2000. Fruit and vegetable crops research report, p. 9. Univ. Ky., Lexington, Agr. Expt. Sta. Bul PR-436.

Sanders, D.C. (ed.). 1999. 2001–2002 vegetable crops guidelines for the southeastern U.S. Vance Publishing., Lincolnshire, Ill.

SAS. 1999. SAS/STAT user's guide, version 8. Cary, N.C.

Saville, D.J. 1990. Multiple comparison procedures: The practical solution. Amer. Stat. 44(2):174–180.

Schultheis, J.R. and S.A. Walters. 1998. Yield and virus resistance of summer squash varieties and breeding lines in North Carolina. HortTechnology 8(1):31–39.

Simonne, E.H. (ed.). 1996a. Fall 1995 commercial vegetable variety trials. Auburn Univ., Ala. Agr. Expt. Sta. Prog. Rpt. 129.

Simonne, E.H. (ed.). 1996b. Spring 1996 commercial vegetable variety trials. Auburn Univ., Ala. Agr. Expt. Sta. Prog. Rpt. Rpt. 130.

Simonne, A.H., E.H. Simonne, R.R. Eitenmiller, H.A. Mills, and N.R. Green. 1997. Ascorbic acid and provitamin A contents in unusually colored bell peppers (*Capsicum annuum* L.) J. Food Comp. Anal. 10(4):299–311.

Simonne, E.H. (ed.). 1997a. Fall 1996 commercial vegetable variety trials. Auburn Univ., Ala. Agr. Expt. Sta. Prog. Rpt. 131.

Simonne, E.H. (ed.). 1997b. Spring 1997 commercial vegetable variety trials. Auburn Univ., Ala. Agr. Expt. Sta. Prog. Rpt. 132.

Simonne, E.H. (ed.). 1998a. Fall 1997 commercial vegetable variety trials. Auburn Univ., Ala. Agr. Expt. Sta. Prog. Rpt. 133.

Simonne, E.H. (ed.). 1998b. Spring 1999 commercial vegetable variety trials. Auburn Univ., Ala., Reg. Bul. 1.

Simonne, E.H., A.H. Simonne and R. Boozer. 1999. Yield, ear characteristics and consumer acceptance of selected white sweet corn varieties in the Southeast. HortTechnology 9(2):289–293.

Simonne, E.H. (ed.). 1999a. Fall 1998 commercial vegetable variety trials. Auburn Univ., Ala., Reg. Bul. 2.

Simonne, E.H. (ed.). 1999b. Spring 1999 commercial vegetable variety trials. Auburn Univ., Ala., Reg. Bul. 3.

Simonne, E.H. (ed.). 2000. Fall 1999 commercial vegetable variety trials. Auburn Univ., Ala., Reg. Bul. 4.

Simonne, E.H., R.T. Boozer, R.N. Baktiyarova, and E.L. Vinson, III. 2000. From artichoke to zucchini: Vegetable varieties for the Southeast. Auburn Univ., Ala. Agr. Expt. Sta. Bul. 640.

Southwick, S.M., J.T. Yeager, J. Osgood, W. Olson, M. Norton, and R. Buchner. 1999. Performance of New Marianna rootstocks in California for 'French' prune. HortTechnology 9(3):498–505.

Stoffella, P.J. and D.N. Maynard. 1988. Stand deficiencies and replanting effects on tomato fruit yield and size. J. Amer. Soc. Hort. Sci. 113:689–693.

Sullivan, J.G. and F.A. Bliss. 1981. Compensation for the missing plants in field experiments with the common bean. HortScience 16(2):185–186.

Swallow, W.H. 1984. Those overworked and oft-misused mean separation procedures-Duncan's, LSD, etc. Plant Dis. 68:919–921.

Tukey, J.W. 1991. The philosophy of multiple comparisons. Stat. Sci. 6(1):100–116.

Wang, M and I.L. Goldman. 1996. Phenotypic variation in free folic acid content among F_1 hybrids an open-pollinated varieties of red beet. J. Amer. Soc. Hort. Sci. 121(6):1040–1042.

Wehner,T.C. 1987. Efficient methods for testing vegetable cultivars. HortScience 22(6):1220–1223.

Wehner, T.C. and C.V. Shaw. 1994. Presentation and analysis of variance results and graphical data. HortScience 29:608.

Westfall, P.H., R.D. Tobias, D. Rom, R.D. Wolfinger and Y. Hochberg. 1999. Multiple comparisons and multiple tests using the SAS system. SAS Institute Inc., Cary, N.C. 5 Apr. 2002. <http://ftp.sas.com/samples/A56648>.