Analysis of Trials with Complex Treatment Structure Using Multiple Contrast Tests

Frank Schaarschmidt¹ and Lea Vaas

Institute of Biostatistics, Natural Sciences Faculty, Leibniz Universität Hannover, Herrenhäuser Str. 2, Hannover D-30419, Germany

Additional index words. augmented factorial, fixed-dose combination, multiple comparisons, simultaneous confidence intervals, linear model

Abstract. Experiments with complex treatment structures are not uncommon in horticultural research. For example, in augmented factorial designs, one control treatment is added to a full factorial arrangement, or an experiment might be arranged as a two-factorial design with some groups omitted because they are practically not of interest. Several statistical procedures have been proposed to analyze such designs. Suitable linear models followed by F-tests provide only global inference for main effects and their interactions. Orthogonal contrasts are demanding to formulate and cannot always reflect all experimental questions underlying the design. Finally, simple mean comparisons following global F-tests do not control the overall error rate of the experiment in the strong sense. In this article, we show how multiple contrast tests can be used as a tool to address the experimental questions underlying complex designs while preserving the overall error rate of the conclusions. Using simultaneous confidence intervals allows for displaying the direction, magnitude, and relevance of the mean comparisons of interest. Along with application in statistical software, shown by two examples, we discuss the possibilities and limitations of the proposed approach.

In agricultural and horticultural research, controlled experiments are set up to evaluate the effect of several treatments and their interactions on physiological or developmental variables. If the levels of a first factor of potential influence are combined with all levels of a second factor, the resulting experiment has a two-factorial, completely cross-classified treatment structure. However, often the experimental questions are manifold and, together with the background knowledge on the practical problem, lead to complex treatment structures. Augmented factorial designs are a common example, in which one or several control treatments are added to a completely crossclassified factorial design. Then, comparisons between the control treatment and the full factorial part are of interest as well as the analysis of the factors arranged in the full factorial part. Another reasonable setting arises when the effect of two factors and their interaction is investigated, but not all combinations between the levels of the two factors are of practical interest. Then, these treatments are reasonably omitted from the experiment, leading to a treatment structure that is crossclassified with a few missing cells.

Among others, Marini (2003) discussed different strategies for analyzing augmented factorial designs. He considered four differ-

ent linear models followed by comparisons of particular treatments. First, a one-way model was performed followed by all pairwise comparisons of the seven treatment means using the least significant difference (LSD) procedure. However, even when protected by a preceding global F-test, decisions based on LSD do not control the familywise error in the strong sense (Hochberg and Tamhane, 1987). Second, a two-way model was performed with subsequent F-tests for the main effect of formulation, the main effect of the concentration, and their interaction. When this approach is used, least square means cannot be estimated properly. Third, a pseudo one-way model comprising all seven treatment groups was used to compute six orthogonal contrasts defining some hypotheses of interest. The orthogonal contrasts involved the comparison of the control group to the average of all other treatments, two contrasts for the comparisons of the three formulations, one contrast for the comparison of the two formulations, and two contrasts for certain interaction effects. However, this approach has severe restrictions; the number of hypotheses that can be formulated is restricted to $k-1$ if k is the number of treatment groups (Dean and Voss, 1999; Marini, 2003; Petersen, 1994). Often, not all hypotheses of practical interest can be included in this type of analysis. Moreover, the problem of multiple comparisons is not taken into account. In a last approach, a mixed model was performed including the concentrations as a quantitative variable and the formulations as a qualitative variable, whereas the blocks were included as random effect.

Piepho et al. (2006) propose another strategy. The authors show how to develop linear models by insertion of an additional variable containing the value CON for observations in the control and the value TREAT for observations belonging to treatments in the complete factorial structure; thus, the new variable represents the fragmentation of the control group and the six treatments and one is able to write out a nested model representing the experimental layout correctly. Note that in this approach, the nesting of effects is not used to represent nested random effects in a hierarchical model. Rather, the nesting operator is used to define an appropriate design matrix for fixed effects. Based on this model, least square means can be estimated properly and an analysis of variance can be performed that takes all existing groups into account and provides F-tests for relevant hypotheses in the complex treatment structure. One disadvantage of this approach is that results of F-tests generally provide only global information about main effects and interaction effects, i.e., a significant result gives evidence for a difference in means among any of the considered treatments. Information about the location of the difference(s), the effect size, or comparisons of particular interest is not available from this approach.

If the interesting experimental questions can be expressed best as a set of comparisons among particular treatment means, a multiple comparison problem results. If testing an increasing number of hypotheses with the number of true hypotheses unknown, the probability of at least one wrong testing decision increases. If it is the aim of the statistical analysis to control the probability of at least one false rejection among all the tested null hypotheses, procedures are needed that control the familywise error (FWE). Computationally simple procedures like the Bonferroni or Scheffé adjustment (e.g., in Nelson, 1989) are suitable for any type of comparisons between means but are known to be conservative because they ignore the correlations among the comparisons. More advanced standard approaches to control the FWE are the tests according to Tukey (e.g., in Hochberg and Tamhane, 1987, citing Tukey, 1953) or Dunnett (1955). However, these do usually not reflect the experimental questions underlying designs with complex treatment structures. Tukey's procedure is appropriate for all pairwise comparisons and therefore is often considered as conservative when testing and adjusting for more hypotheses than are actually of interest. Dunnett's procedure performs comparisons of several treatments to a control. With complex treatment structures, usually more than these comparisons are of interest. In the recent years, multiple comparisons procedures have been made available (Bretz et al., 2001, 2002; Hothorn et al., 2008a; Westfall 1997; Westfall et al., 1999), which provide the feature of controlling the FWE for a certain user-defined set of comparisons formulated as multiple contrasts of the treatment means. The number of comparisons as well as the correlation among

Received for publication 4 June 2008. Accepted for publication 23 Oct. 2008.

We thank Prof. L.A. Hothorn, M. Hasler, and the referees for their helpful comments on earlier versions of this paper.

¹To whom reprint requests should be addressed; e-mail schaarschmidt@biostat.uni-hannover.de.

the comparisons is taken into account specifically for each user-defined set of contrasts. Although these methods are computationally more demanding than well-known standard procedures, they are numerically available in various softwares.

If multiple hypotheses are tested without controlling the FWE rate at α , the probability of finding at least one of the considered differences to be significant when this difference is in fact zero can be markedly higher than α . For motivation of multiple comparison procedures, we present simulated FWE probabilities in Tables 1 and 2.

Assuming that no true effect is present in an experiment with three, five, or 10 groups, Table 1 shows the probabilities to find no significant difference and one, two, or more significant differences when all pairwise comparisons are performed among these treatments without multiplicity adjustment or protection by a global F-test. When there are only three treatments, resulting in three tests, one will conclude for at least one significant difference in ≈ 12 of 100 experiments. If 10 treatments are compared in such a way, ≈ 61 of 100 experiments will identify an effect that is not reproducible in a follow-up experiment because a relatively large difference in means simply occurred by chance.

Table 2 shows similar numbers for the situation that two, four, or nine independent two-sided t tests are performed. This is a situation close to that of testing two, four, or nine orthogonal (independent) contrasts with a common residual degree of freedom (df). With only two independent tests, the overall chance of a Type I error is 0.0975, i.e., $1 - (1 - 0.05)^2$, whereas when testing nine independent hypotheses, the odds are $\approx 1:2$ to observe an effect that is not present in truth. When overall (familywise) Type I errors of this magnitude are not acceptable, LSD or single df contrasts should not be used, and multiple contrast procedures, as described in the following, are more appropriate.

In the following section, we introduce two example data sets and briefly review the well-

known concept of multiple contrast tests. Subsequently, we show the application of multiple contrast tests to the two examples and give an interpretation of the results. Finally, we discuss the advantages and limitations of the proposed approach.

Material and Methods

Two example data sets. Marini (2003) describes the analysis of an experiment with an augmented factorial treatment structure. The effect of three formulations of gibberellic acid (f1, f2, f3), each in two different concentrations, 10 and 20 mg·L⁻¹, on fruit set of apple was investigated. Additional to this two-factorial structure, a control group (application of water) was included in the design. The composition of the treatments is summarized in Table 3.

The seven treatment groups were arranged in a randomized complete block design using six apple trees as blocks. As the response variable, the ratio of fruits (65 d after bloom) per 100 flower clusters was presented (Marini, 2003). In previous analyses, Marini (2003) and Piepho et al. (2006) proposed the following experimental questions; the comparison of all treatments with the untreated control aims to show that the experiment was sensitive to reveal marked effects on the fruit set. The comparison of increasing dosages pooled over the formulations aims to assess whether and until which concentration a dose effect is present. For the two-factorial part of the trial, interest was also in the main effects of the formulations and concentrations taking the possibility of an interaction into account. By the formulation of orthogonal contrasts, Marini (2003) addressed the question for interactions more explicitly; namely, whether the different formulations affect the increase of fruit set from dose 10 to dose 20.

As a second example, we consider a fixed-dose combination experiment originally published by Adeli and Varco (2002). The effects of potassium (K) management on cotton yield (kg·ha⁻¹) were investigated. The objective

Table 3. Experimental layout of Example 1 with three gibberellic acid formulations (f1, f2, f3), each in two different concentrations (10 and 20 mg·L⁻¹), and a water control (H₂O).

Concn	Formulation			
	H ₂ O	f1	f2	f3
0	CON	—	—	—
10	—	f1(10)	f2(10)	f3(10)
20	—	f1(20)	f2(20)	f3(20)

was “to determine potassium (K) fertilizer rate and placement effects on cotton lint yield” (Adeli and Varco, 2002). The experimental setup contained two different application methods for K: “broadcast” (Bc) and “banded” (Bn), each in four concentrations. The factorial arrangement included 0, 68, and 136 kg·ha⁻¹ K broadcast in all possible combinations with 0, 34, and 68 kg·ha⁻¹ K banded application. Two additional treatments, with 204 kg·ha⁻¹ K broadcast with zero banded and 102 kg·ha⁻¹ K banded with zero broadcast application, were included in the design. This treatment structure can be imagined as arisen from a complete two-factorial structure with those cells omitted that lead to inappropriately high total dosages of kg·ha⁻¹ K. Table 4 summarizes the treatment combinations. The doses were chosen carefully so that the total amount of K is the same in four pairs of treatments (in Table 4, members of each pair are given the same symbol). The resulting 11 treatment groups were assumed to be arranged in a completely randomized design with replication number $n = 12$. The data used for the analysis are simulated based on the published summary statistics.

The analysis should give information about supposed beneficial influences of the different placement methods and an expected diverging effect of the two methods. Furthermore, the aim is to find treatment combinations resulting in superior cotton lint yield given that the total K fertilizer application is the same. Note that regression method, and more specifically response surface regression, is a viable option to analyze this example. Compared with the methods discussed subsequently, regression methods have the advantage of using fewer parameters to describe the data, but additionally rely on assumptions concerning the dose-response relationship. Response surface regression is particularly useful when the aim is to estimate the optimum combination of the two quantitative variables. For an introduction to response surface regression, see, for example, Montgomery (2005).

Simultaneous confidence intervals for user-defined multiple contrasts. In this section, we review the concept of multiple contrast tests as, for example, described in Bretz et al. (2001). A simple linear model to explain the observations Y_{ij} is:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

with Y_{ij} denoting the j th observation of the i th treatment group, with $i = 1, \dots, k$, and

Table 1. Probability to reject x out of M tested hypotheses when there is no difference among three, five, and 10 treatments and all pairwise comparisons are performed using multiple t tests (comparisons wise Type I error probability 5%) without adjustment for multiple testing or the protection of a global F-test.^z

Number of treatments	Number of hypotheses M	Probability to reject H_0 for x hypotheses			
		$x = 0$	$x = 1$	$x = 2$	$x \geq 3$
3	3	0.881	0.090	0.029	0.000
5	10	0.723	0.135	0.083	0.060
10	45	0.390	0.139	0.121	0.350

^zThe probabilities are estimated by 10,000 simulation runs for each setting.

Table 2. Probability to reject x out of M tested hypotheses when there is no true difference in two, four, and nine independent two-sided contrasts (comparisons wise Type I error probability 5%) without adjustment for multiple testing.^z

Number of treatments	Number of hypotheses M	Probability to reject H_0 for x hypotheses			
		$x = 0$	$x = 1$	$x = 2$	$x \geq 3$
3	2	0.903	0.094	0.003	—
5	4	0.819	0.164	0.016	0.002
10	9	0.644	0.284	0.063	0.009

^zThe probabilities are estimated by 10,000 simulation runs for each setting.

Table 4. Treatment structure of Example 2 (Adeli and Varco, 2002) comprising 11 combinations of broadcast (Bc) and banded (Bn) application of different doses of potassium (K) fertilizer.^z

Banded application	Broadcast application			
	0 kg·ha ⁻¹ K	68 kg·ha ⁻¹ K	136 kg·ha ⁻¹ K	204 kg·ha ⁻¹ K
0 kg·ha ⁻¹ K	Bc0Bn0	Bc68Bn0 *	Bc136Bn0 #	Bc204Bn0 °
34 kg·ha ⁻¹ K	Bc0Bn34	Bc68Bn34 +	Bc136Bn34	—
68 kg·ha ⁻¹ K	Bc0Bn68 *	Bc68Bn68 #	Bc136Bn68 °	—
102 kg·ha ⁻¹ K	Bc0Bn102 +	—	—	—

^zTreatments resulting in the same total amount of kg·ha⁻¹ K are labeled with common symbols, * (68 kg·ha⁻¹ K), + (102 kg·ha⁻¹ K), # (136 kg·ha⁻¹ K), and ° (204 kg·ha⁻¹ K).

$$j = 1, \dots, n_i$$

μ_i denoting the mean of the i th treatment group; and

ε_{ij} denoting the residual error for the j th observation in the i th group.

The errors ε_{ij} are assumed to be independent, i.e., the observations are derived from a completely randomized design and to be Gaussian distributed with equal variances: $\varepsilon_{ij} \sim N(0, \sigma^2)$. From fitting this one-way model, we derive estimates $\hat{\mu}_i$ for the treatment means and $\hat{\sigma}^2$ for the residual error. The questions of interest can then be stated as differences of the treatment means μ_i or, more generally, as contrasts of the treatment means μ_i . A contrast L is a weighted difference of the $\mu_1, \mu_2, \dots, \mu_k$, where the weights c_i are chosen such that a certain difference of interest is built: $L = \sum c_i \mu_i = c_1 \mu_1 + c_2 \mu_2 + \dots + c_k \mu_k$. For example, in a design comprising four treatments, $i = 1, 2, 3, 4$, the difference between Treatments 1 and 2, $\mu_2 - \mu_1$ can be written as $\sum c_i \mu_i$ with $c_1 = -1, c_2 = 1, c_3 = 0, c_4 = 0$. However, also, more complicated differences, like for example the difference of the first treatment's mean to the average mean of the three remaining treatments, $(\mu_2 + \mu_3 + \mu_4)/3 - \mu_1$ can be formulated: $c_1 = -1, c_2 = 1/3, c_3 = 1/3, c_4 = 1/3$.

For the choice of the c_i s, we impose only the restriction that the sum of c_i s should be zero, $\sum c_i = 0$. This ensures that the contrast has expectation 0 if in fact all μ_i are equal. Moreover, we usually choose the c_i such that the sum of all negative coefficients is -1 and hence the sum of all positive coefficients is 1. Then, we can interpret the confidence intervals for the contrasts as differences of (weighted averages of) treatment means. Usually, several, say M , such contrasts are necessary to represent the experimental questions of interest. Note that there are no further restrictions on the choice of the c_i depending on the remaining set of contrasts. That is, there is no necessity to define the M contrasts orthogonal to each other, and there is no restriction on the number of contrasts M , as in the case of orthogonal single df contrasts. The test statistic for one contrast can be calculated from:

$$T = \frac{\sum_{i=1}^k c_i \hat{\mu}_i}{\hat{\sigma} \sqrt{\sum_{i=1}^k \frac{c_i^2}{n_i}}}$$

Simultaneous confidence intervals for the M contrasts can be calculated from

$$\left[\sum_{i=1}^k c_i \hat{\mu}_i \pm q_{1-\alpha, M, R}^{two-sided} \hat{\sigma} \sqrt{\sum_{i=1}^k \frac{c_i^2}{n_i}} \right],$$

using $q_{1-\alpha, M, R}^{two-sided}$ as the critical value calculated from the multivariate t-distribution with dimension M and correlation matrix R . For general contrasts, the correlation matrix R has a complicated structure with elements depending on the sample sizes n_i and the contrast coefficients c_i .

For a particular contrast, the null hypothesis that the difference defined by the contrast has the value zero can be rejected if $|T| > q_{1-\alpha, M, R}^{two-sided}$ or if the confidence interval does not contain the value zero. How such critical values can be obtained is described by Westfall et al. (1999) or Bretz et al. (2001). The presented confidence intervals can be obtained using the LSMESTIMATE statement in the SAS PROC GLIMMIX (SAS Institute, 2006) or the package multcomp (Bretz et al., 2002; Hothorn et al., 2008b) in the free statistical software R (R Development Core Team, 2008). In the remaining part of the article, R-2.6.2 is used with multcomp, Version 0.993-2.

We favor the graphical display of the simultaneous confidence intervals for reporting the results of a statistical analysis. From such plots, the significance of a particular difference at a FWE level α can be inferred if the value zero is not included in the confidence interval. Additionally, the direction (decrease or increase), magnitude, and, possibly, relevance of an effect can be assessed. If interpreting the relevance of the measured effect is of interest, confidence intervals are advantageous compared with P values because of displaying the effect size in the scale of the measured variable rather than in the scale of probability. Finally, the uncertainty concerning

the estimated effect, depending on the sample variance and the sample size, is displayed by the width of the confidence interval.

Results

Evaluation of Example 1. From previous discussions of Example 1 (Marini, 2003; Piepho et al., 2006), the following experimental questions can be deduced: Is the experimental setting capable of revealing effects on the response? Do the formulations differ? Do the concentrations differ?

In the following, hypotheses that might be of interest are stated as differences of treatment means using the acronyms introduced in Table 3. The contrast coefficients (c_i) leading to the stated differences are summarized in Table 5. The first difference of interest compares the pooled means of all treatments versus the untreated control group:

$$1. \mu_{\text{all treatments}} - \mu_{\text{CON}} = 0.$$

For the detection of main effects of the formulation, three pairwise comparisons could be of interest, which pool over the two concentrations:

$$2. \mu_{f2} - \mu_{f1} = 0;$$

$$3. \mu_{f3} - \mu_{f1} = 0;$$

$$4. \mu_{f3} - \mu_{f2} = 0.$$

Analogously, the comparison of the two concentrations can be done by pooling over the formulations:

$$5. \mu_{(20)} - \mu_{(10)} = 0.$$

Interactions can be detected by building differences of differences (Petersen, 1994), i.e., comparing the difference between concentrations 10 and 20 between Formulations 1 and 2, and so on, as is done in Comparisons 6 through 8 in Table 5.

$$6. [\mu_{f1(20)} - \mu_{f1(10)}] - [\mu_{f2(20)} - \mu_{f2(10)}] = 0$$

$$7. [\mu_{f1(20)} - \mu_{f1(10)}] - [\mu_{f3(20)} - \mu_{f3(10)}] = 0$$

$$8. [\mu_{f2(20)} - \mu_{f2(10)}] - [\mu_{f3(20)} - \mu_{f3(10)}] = 0$$

Note that the stated contrasts reflect similar hypotheses as have been tested using four

Table 5. Contrast coefficients (c_i) are summarized for the multiple contrast tests indicated in the above text.^z

Number	Comparison	Treatment:					
		Control	f1(10)	f1(20)	f2(10)	f2(20)	f3(10)
		Contrast coefficients:					
1	All treatments-CON	1	-1/6	-1/6	-1/6	-1/6	-1/6
2	f2 - f1	0	-1/2	-1/2	1/2	1/2	0
3	f3 - f1	0	-1/2	-1/2	0	0	1/2
4	f3 - f2	0	0	0	-1/2	-1/2	1/2
5	(20) - (10)	0	-1/3	1/3	-1/3	1/3	-1/3
6	Interaction f1 versus f2	0	-1	1	1	-1	0
7	Interaction f1 versus f3	0	-1	1	0	0	1
8	Interaction f2 versus f3	0	0	0	-1	1	-1

^zCalculated are the comparison of control versus the pooled treatments (the first contrast), the comparisons of the formulations pooled over concentrations (Contrasts 2, 3, 4), the comparison of the concentrations pooled over formulations (Contrast 5), and contrasts for the interactions between formulation and concentration (Contrasts 6, 7, 8).

F-tests after a suitable model reformulation in the analysis by Piepho et al. (2006).

Simultaneous confidence intervals for the comparisons formally defined in Table 5 are plotted in Figure 1. The complete analysis has been performed in one simple procedure with all done tests adjusted for multiplicity inherently.

Altogether, the six gibberellin treatments lead to a significant increase in the number of fruits per flower cluster. Hence, the experimental setting is capable of revealing effects of the gibberellin treatments compared with an untreated control. On average, over the six gibberellin treatments, we can expect an increase of at least five fruits per 100 clusters more than in the untreated control with 95% confidence. Practically, the mean increase in the response is not very interesting; it is not possible to decide which treatments mainly contribute to the overall effect. Moreover, controlling the FWE for all eight hypotheses, none of the remaining tests is significant at the 5% level. None of the differences among the three formulations, each pooled over the concentrations, are significantly different from 0. Although comparing the average effect of the three formulations at concentration 20 with their average effect at concentration 10 shows a mean increase, the observed difference is not significant when controlling the overall Type I error probability at 5%. Finally, none of the three interaction contrasts differs significantly from zero, although the mean increase in the response when increasing the concentration from 10 to 20 is somewhat more pronounced in f1 and f2 compared with f3 (Comparisons 7 and 8). That is, given the limited sample size of the trial, we cannot conclude that there are differences among the formulations, between the concentrations, and cannot prove the presence of interactions when controlling the FWE at 5%.

The contrasts in Table 5 were constructed to show that hypotheses similar to those of the analysis of variance F-tests used by Piepho et al. (2006) can be tested using a multiple contrast approach. In practice, other comparisons can be more interesting and are as simple to implement. In the following, we show an analysis, alternative to that in Table 5.

First, it could be of interest whether any of the six gibberellin treatments leads to a change in the number of fruit limbs per number of flowers. Hypotheses 1 to 6 represent these comparisons; the resulting contrast coefficients are presented in Table 6.

1. $\mu_{f1(10)} - \mu_{CON} = 0;$
2. $\mu_{f1(20)} - \mu_{CON} = 0;$
3. $\mu_{f2(10)} - \mu_{CON} = 0;$
4. $\mu_{f2(20)} - \mu_{CON} = 0;$
5. $\mu_{f3(10)} - \mu_{CON} = 0;$
6. $\mu_{f3(20)} - \mu_{CON} = 0.$

Second, it might be of interest whether the formulations differ, taking the possibility of an interaction into account. This may result in comparisons of the different formulations at each of the two concentration levels:

7. $\mu_{f2(10)} - \mu_{f1(10)} = 0;$
8. $\mu_{f3(10)} - \mu_{f1(10)} = 0;$
9. $\mu_{f3(10)} - \mu_{f2(10)} = 0;$
10. $\mu_{f2(20)} - \mu_{f1(20)} = 0;$
11. $\mu_{f3(20)} - \mu_{f1(20)} = 0;$
12. $\mu_{f3(20)} - \mu_{f2(20)} = 0.$

Finally, the concentrations 10 and 20 could be compared separately for each formulation:

13. $\mu_{f1(20)} - \mu_{f1(10)} = 0;$
14. $\mu_{f2(20)} - \mu_{f2(10)} = 0;$
15. $\mu_{f3(20)} - \mu_{f3(10)} = 0.$

These are 15 comparisons in total, which could not have been performed using orthogonal contrasts. The correlation structure among these 15 comparisons is not trivial; however, it is taken into account inherently by the statistical software.

Figure 2 shows simultaneous 95% confidence intervals for the contrasts defined in Table 6.

This evaluation results in a more informative interpretation than the first approach: Two of the six gibberellin treatments result in a significant increase of the number of fruits per 100 flower clusters. With 95% confidence, we can expect an increase in the mean number of fruits per 100 flower clusters of at least two when Formulation 1 with Concentration 20 is applied. Using Formulation 2 with Concentration 20, one can expect at least 18 fruits per 100 clusters more than in the untreated control. The remaining combinations of formulation and concentration led to a mean increase of the number of fruits per cluster, but, controlling the FWE for all comparisons, the observed differences are not significant. Furthermore, the pairwise differences among the formulations are not significant when considered separately for each concentration (Comparisons 7 through 12). Finally, a difference between Concentrations 10 and 20 cannot be shown at the 5% level for any of the three formulations.

Evaluation of Example 2. The experiment presented by Adeli and Varco (2002) shows a more complex treatment structure. First, it could be of interest which K rate or application method increases the yield compared with control. That is, the differences of the 10 different K treatments to the untreated control treatment Bc0Bn0 are of primary practical interest. Furthermore, interest is in the magnitude of increase that can at least be expected with high probability, i.e., in lower confidence limits. Hypotheses 1 to 3 compare the three treatments with only broadcast application with the untreated control group:

1. $\mu_{Bc68Bn0} - \mu_{Bc0Bn0} = 0;$
2. $\mu_{Bc136Bn0} - \mu_{Bc0Bn0} = 0;$
3. $\mu_{Bc204Bn0} - \mu_{Bc0Bn0} = 0.$

Hypotheses 4, 5, and 6 compare the treatments with only banded application with the untreated control group:

4. $\mu_{Bc0Bn34} - \mu_{Bc0Bn0} = 0;$
5. $\mu_{Bc0Bn68} - \mu_{Bc0Bn0} = 0;$
6. $\mu_{Bc0Bn102} - \mu_{Bc0Bn0} = 0.$

Hypotheses 7 to 10 compare the treatments with mixed application of banded and broadcast application with the untreated control:

Fig. 1. Simultaneous 95% confidence intervals for the eight contrasts formulated in Table 5. Dots mark the point estimates of the differences of interest and parentheses mark the limits of the simultaneous 95% confidence regions for these differences. Calculated are the comparison of control versus the pooled effect of all six treatments (the first interval), the comparison of the formulations with pooled concentrations (Contrasts 2, 3, 4), the comparison of the concentration with pooled formulations (Contrast 5), and the interactions between the formulations (Contrasts 6, 7, 8).

Table 6. Contrast coefficients (c_i) of 15 contrasts for the alternative evaluation of Example 1.^z

		Treatment						
		Control	f1(10)	f1(20)	f2(10)	f2(20)	f3(10)	f3(20)
Number	Comparison	Contrast coefficients						
1	f1(10) – CON	–1	1	0	0	0	0	0
2	f1(20) – CON	–1	0	1	0	0	0	0
3	f2(10) – CON	–1	0	0	1	0	0	0
4	f2(20) – CON	–1	0	0	0	1	0	0
5	f3(10) – CON	–1	0	0	0	0	1	0
6	f3(20) – CON	–1	0	0	0	0	0	1
7	f2(10) – f1(10)	0	–1	0	1	0	0	0
8	f3(10) – f1(10)	0	–1	0	0	0	1	0
9	f3(10) – f2(10)	0	0	0	–1	0	1	0
10	f2(20) – f1(20)	0	0	–1	0	1	0	0
11	f3(20) – f1(20)	0	0	–1	0	0	0	1
12	f3(20) – f2(20)	0	0	0	0	–1	0	1
13	f1(20) – f1(10)	0	–1	1	0	0	0	0
14	f2(20) – f2(10)	0	0	0	–1	1	0	0
15	f3(20) – f3(10)	0	0	0	0	0	–1	1

^zDefined are the comparisons of the six gibberellin treatments versus the control (Contrasts 1–6), the comparisons of the formulations separate for each concentration (Contrasts 7–12), and the comparisons of Concentrations 10 and 20 separate for each formulation.

Fig. 2. Simultaneous 95% confidence intervals for the 15 contrasts defined in Table 6. Dots mark the point estimates of the differences of interest and parentheses mark the limits of the simultaneous 95% confidence regions for these differences. Calculated are differences of the single treatment means to the mean of the control (Contrasts 1–6), the comparisons of the formulations separate for each concentration (Contrasts 7–12), and the comparisons between Concentrations 10 and 20 separate for each formulation (Contrasts 13–15).

$$7. \mu_{\text{Bc68Bn34}} - \mu_{\text{Bc0Bn0}} = 0;$$

$$8. \mu_{\text{Bc68Bn68}} - \mu_{\text{Bc0Bn0}} = 0;$$

$$9. \mu_{\text{Bc136Bn34}} - \mu_{\text{Bc0Bn0}} = 0;$$

$$10. \mu_{\text{Bc136Bn68}} - \mu_{\text{Bc0Bn0}} = 0.$$

Furthermore, the aim could be to analyze whether any treatment that combines broadcast with banded application leads to an increase in yield compared with treatments resulting in the same amount of K with only one application method, broadcast or banded. The corresponding comparisons are:

$$11. \mu_{\text{Bc68Bn34}} - \mu_{\text{Bc0Bn102}} = 0;$$

$$12. \mu_{\text{Bc68Bn68}} - \mu_{\text{Bc136Bn0}} = 0;$$

$$13. \mu_{\text{Bc136Bn68}} - \mu_{\text{Bc204Bn0}} = 0.$$

Finally, it could be of interest whether banded application of 68 kg-ha⁻¹ K leads to

an increase in yield compared with broadcast application of the same amount:

$$14. \mu_{\text{Bc0Bn68}} - \mu_{\text{Bc68Bn0}} = 0.$$

The contrast coefficients resulting from these 14 contrasts are presented in Table 7.

Simultaneous confidence intervals for the contrasts defined in Table 7 are plotted in Figure 3.

With 95% confidence, one can state that broadcast application of K leads to a significant increase in yield when applied with 136 kg-ha⁻¹ or 204 kg-ha⁻¹. For banded application alone, no significant effect can be found. All combinations of banded and broadcast applications lead to a significant increase of cotton yield compared with the untreated control. Applying the combinations Bc68Bn34, Bc68Bn68, and Bc136Bn34 leads to a mean increase in yield compared with the untreated control of at least 30 kg-ha⁻¹, 55 kg-ha⁻¹, and 180 kg-ha⁻¹, respectively. None of the treatments combining broadcast and banded appli-

cation leads to a significant increase in yield compared with treatments applying the same amount of K with only one application method. Finally, there is no significant difference in yield between banded and broadcast application of 68 kg-ha⁻¹ K.

Discussion

This article shows that simultaneous confidence intervals for multiple contrasts are a flexible method to evaluate factorial experiments with nonstandard treatment structures as, for example, augmented factorial designs or experiments with two or more factors, which are crossclassified with some factor combinations omitted. The strategy can be summarized as follows; estimators for the treatment means and variance are derived from a simple general linear model with all treatments combined in a single factor (i.e., a pseudo-one-way layout or cell means model). Contrast coefficients are chosen by the user such that the hypotheses of interest are reflected as differences of (weighted averages of) treatment means. Like other procedures following the general linear model (Marini, 2003; Piepho et al., 2006), the described procedure relies on the assumptions that the observations are mutually independent and continuous with normal distributed errors and homogeneous variances. The method is computationally available for the R environment for statistical computing as well as in SAS.

Compared with other methods that have been proposed for evaluation of experiments with complex treatment structures, the described method has a number of advantages. The individual contrasts give more specific information than the global decisions provided by an analysis of variance F-test. When simultaneous confidence intervals are used, the significance, relevance, and direction (increase or decrease) of the effect of interest as well as the uncertainty concerning the estimates can be interpreted in a scale close to that of the measured variable, which is often easier than interpreting *P* values in the scale of probability. Compared with orthogonal single df contrasts, the contrasts formulated for the method described in this article do not need to be mutually orthogonal and are not restricted in their number. Finally, the overall Type I error probability is controlled inherently for a user-defined set of contrasts.

The described method is of limited use if experiments are analyzed that comprise many, say more than 20, treatments. Being still methodologically correct, it then has the drawback that the contrast matrix becomes huge and it is hard to control for typos in the definition of comparisons of interest. Also, when the number of contrasts becomes very high, computations can be very time-consuming or impossible. For some scenarios, the methods described by Piepho et al. (2006) can then be more appropriate.

In this article, we discuss only complex treatment structures and for brevity assume a simple randomization structure

Table 7. Contrast coefficients (c_i) for the 14 comparisons among the 11 treatments of the cotton example.^z

K application				Treatment combinations								
	Bc	0	68	136	204	0	0	0	68	68	136	136
	Bn	0	0	0	0	34	68	102	34	68	34	68
	Total K	0	68	136	204	34	68	102	102	136	170	204
Number	Comparison	Contrast coefficients										
1	Bc68Bn0 – Bc0Bn0	-1	1	0	0	0	0	0	0	0	0	0
2	Bc136Bn0 – Bc0Bn0	-1	0	1	0	0	0	0	0	0	0	0
3	Bc204Bn0 – Bc0Bn0	-1	0	0	1	0	0	0	0	0	0	0
4	Bc0Bn34 – Bc0Bn0	-1	0	0	0	1	0	0	0	0	0	0
5	Bc0Bn68 – Bc0Bn0	-1	0	0	0	0	1	0	0	0	0	0
6	Bc0Bn102 – Bc0Bn0	-1	0	0	0	0	0	1	0	0	0	0
7	Bc68Bn34 – Bc0Bn0	-1	0	0	0	0	0	0	1	0	0	0
8	Bc68Bn68 – Bc0Bn0	-1	0	0	0	0	0	0	0	1	0	0
9	Bc136Bn34 – Bc0Bn0	-1	0	0	0	0	0	0	0	0	1	0
10	Bc136Bn68 – Bc0Bn0	-1	0	0	0	0	0	0	0	0	0	1
11	Bc68Bn34 – Bc0Bn102	0	0	0	0	0	0	-1	1	0	0	0
12	Bc68Bn68 – Bc136Bn0	0	0	-1	0	0	0	0	0	1	0	0
13	Bc136Bn68 – Bc204Bn0	0	0	0	-1	0	0	0	0	0	0	1
14	Bc0Bn68 – Bc68Bn0	0	-1	0	0	0	1	0	0	0	0	0

^zCalculated are the comparisons of the 10 different potassium (K) treatments to the untreated the control (Contrasts 1–10) and the comparisons of groups with equal amount of K but different application methods (Contrasts 11–14).

Fig. 3. Simultaneous 95% lower confidence limits for the 14 contrasts defined in Table 7. Dots mark the point estimates of the differences of interest and parentheses mark the limits of the simultaneous 95% confidence regions for these differences. Calculated are the comparison of the control versus each of the single fertilizer combinations (1–10) and the comparisons of groups with equal amount of potassium but different application methods (11–14).

and homoscedastic Gaussian error distribution for the response variable. Nevertheless, the concept of multiple contrast tests can be extended so that situations with different assumptions or randomization schemes are also covered. Block effects or more complex randomization structures may be included as random effects in a linear mixed effects model, whereas the complex treatment structure remains in the fixed part (for example Piepho et al., 2003). Computationally, approximate simultaneous confidence intervals for multiple contrasts in mixed models are covered in the SAS PROC GLIMMIX as well as the R package multcomp. When the assumption of the Gaussian distribution is not adequate but counts or proportions are considered, generalized linear models (McCullagh and Nelder, 1989; Piepho, 1999) are an alternative. By default, the primary comparisons are then performed on the log scale for count data and on the logit scale for binomial

proportions. Using the inverse link function to transform back results in confidence intervals for ratios of means and odds ratios when the log and logit link is used, respectively. Again, these cases are computationally solved in PROC GLIMMIX and multcomp. However, also in the general linear model with Gaussian errors, the comparisons of interest could be formulated in terms of ratios rather than in differences of means (Dilba et al., 2006). When interest is in a combination of one-sided and two-sided hypotheses, Braat et al. (2008) provide a method related to the methods shown in this article.

Literature Cited

- Adeli, A. and J.J. Varco. 2002. Potassium management effects on cotton yield, nutrition, and soil potassium level. *J. Plant Nutr.* 25:2229–2242.
- Braat, S., D. Gerhard, and L.A. Hothorn. 2008. Joint one-sided and two-sided simultaneous

confidence intervals. *J. Biopharm. Stat.* 18:293–306.

- Bretz, F., A. Genz, and L.A. Hothorn. 2001. On the numerical availability of multiple comparison procedures. *Biom. J.* 43:645–656.
- Bretz, F., T. Hothorn, and P.H. Westfall. 2002. On multiple comparisons in R. *R News* 2:14–17.
- Dean, A. and D. Voss. 1999. Design and analysis of experiments. Springer-Verlag, New York, NY.
- Dilba, G., F. Bretz, and V. Guiard. 2006. Simultaneous confidence sets and confidence intervals for multiple ratios. *J. Stat. Plan. Infer.* 136:2640–2658.
- Dunnnett, C.W. 1955. A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Stat. Assoc.* 50:1096–1121.
- Hochberg, A.C. and Y. Tamhane. 1987. Multiple comparison procedures. Wiley, New York, NY.
- Hothorn, T., F. Bretz, and P. Westfall. 2008a. Simultaneous inference in general parametric models. *Biometrical Journal* 50:346–363.
- Hothorn, T., F. Bretz, P. Westfall, and R.M. Heiberger. 2008b. Multcomp: Simultaneous inference for general linear hypotheses. R package version 0.993-2.
- Marini, R.P. 2003. Approaches to analyzing experiments with factorial arrangements of treatments plus other treatments. *HortScience* 38:117–120.
- McCullagh, P. and J.A. Nelder. 1989. Generalized linear models. Chapman & Hall/CRC, Boca Raton, FL.
- Montgomery, D.C. 2005. Design and analysis of experiments. 6th Ed. Wiley, Hoboken, NJ.
- Nelson, P.R. 1989. Multiple comparisons of means using simultaneous confidence intervals. *J. Qual. Technol.* 21:232–289.
- Petersen, R.G. 1994. Agricultural field experiments—Design and analysis. Marcel Dekker, New York, NY.
- Piepho, H.-P. 1999. Analysing disease incidence data from designed experiments by generalized linear mixed models. *Plant Pathol.* 48:668–674.
- Piepho, H.-P., A. Büchse, and K. Emrich. 2003. A hitchhiker's guide to mixed models for randomized experiments. *J. Agron. Crop Sci.* 189:310–322.
- Piepho, H.-P., E.R. Williams, and M. Fleck. 2006. A note on the analysis of designed experiments with complex treatment structure. *HortScience* 41:446–452.
- Hochberg, A.C. and Y. Tamhane. 1987. Multiple comparison procedures. Wiley, New York, NY.
- Hothorn, T., F. Bretz, and P. Westfall. 2008a. Simultaneous inference in general parametric models. *Biometrical Journal* 50:346–363.
- Hothorn, T., F. Bretz, P. Westfall, and R.M. Heiberger. 2008b. Multcomp: Simultaneous inference for general linear hypotheses. R package version 0.993-2.
- R Development Core Team. 2008. R: A language and environment for statistical computing. Version 2.6.2. R Foundation for Statistical Computing, Vienna, Austria.
- SAS Institute. 2006. The GLIMMIX procedure, June 2006. SAS Institute, Cary, NC.
- Tukey, J. 1953. The problem of multiple comparisons, unpublished manuscript, reprinted in: Braun, H.I. (Ed.) 1994. The collected works of John W. Tukey. VIII. Multiple comparisons. Chapman and Hall, New York, NY.
- Westfall, P.H. 1997. Multiple testing of general contrasts using logical constraints and correlations. *J. Amer. Stat. Assoc.* 92:299–306.
- 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, Cary, NC.

Appendix

The two example data sets are available at <http://www.biostat.uni-hannover.de/software/>. After loading the data sets into the R workspace under the names ExFruitset (Example 1) and ExKCotton (Example 2), the following R code reproduces the analyses of the two examples shown in this article.

SAS program files, including the data sets and the calculation of the simultaneous intervals, are available at <http://www.biostat.uni-hannover.de/software/>.

The following R code reproduces the calculation of simultaneous confidence intervals plotted in Figure 1:

```
# Define Block as a factor variable and fit the linear model
ExFruitset$Block <- as.factor(ExFruitset$Block)
modell1 <- lm(Fruitset ~ Block + Treatment, data = ExFruitset)
# Load library multcomp
library(multcomp)
# Define the contrast matrix in Table 5
contrast1 <- rbind(
  "1. all treatments - CON"=c(-1,1/6,1/6,1/6,1/6,1/6,1/6),
  "2. f2 - f1"=          c(0,-1/2,-1/2,1/2,1/2,0,0),
  "3. f3 - f1"=          c(0,-1/2,-1/2,0,0,1/2,1/2),
  "4. f3 - f2"=          c(0,0,0,-1/2,-1/2,1/2,1/2),
  "5. (20)-(10)"=        c(0,-1/3,1/3,-1/3,1/3,-1/3,1/3),
  "6. interaction f1 vs f2"=c(0,-1, 1, 1,-1, 0, 0),
  "7. interaction f1 vs f3"=c(0,-1, 1, 0, 0, 1,-1),
  "8. interaction f2 vs f3"=c(0, 0, 0,-1, 1, 1,-1))
# Multiple comparisons and simultaneous confidence intervals
comps1 <- glht(modell1, linfct = mcp(Treatment = contrast1),
  alternative="two.sided")
SCI1<-confint(comps1)
# create the plot in Fig. 1
windows(8,5); par(mar=c(5, 12, 4, 1) + 0.1)
plot(SCI1, main="", xlab="Difference in fruit set
(number of fruits per 100 flower clusters)")
```

The following R code reproduces the calculation of simultaneous confidence intervals plotted in Figure 2:

```
# Define the contrasts in Table 6
contrasts2 <- rbind (
  "1. f1(10) - CON" =      c(-1, 1, 0, 0, 0, 0, 0),
  "2. f1(20) - CON" =      c(-1, 0, 1, 0, 0, 0, 0),
  "3. f2(10) - CON" =      c(-1, 0, 0, 1, 0, 0, 0),
  "4. f2(20) - CON" =      c(-1, 0, 0, 0, 1, 0, 0),
  "5. f3(10) - CON" =      c(-1, 0, 0, 0, 0, 1, 0),
  "6. f3(20) - CON" =      c(-1, 0, 0, 0, 0, 0, 1),
  "7. f2(10) - f1(10)"=    c(0,-1, 0, 1, 0, 0, 0),
  "8. f3(10) - f1(10)"=    c(0,-1, 0, 0, 0, 1, 0),
  "9. f3(10) - f2(10)"=    c(0, 0, 0,-1, 0, 1, 0),
  "10. f2(20) - f1(20)"=   c(0, 0,-1, 0, 1, 0, 0),
  "11. f3(20) - f1(20)"=   c(0, 0,-1, 0, 0, 0, 1),
  "12. f3(20) - f2(20)"=   c(0, 0, 0, 0,-1, 0, 1),
  "13. f1(20) - f1(10)"=   c(0, -1,1, 0, 0, 0, 0),
  "14. f2(20) - f2(10)"=   c(0, 0, 0, -1, 1, 0, 0),
  "15. f3(20) - f3(10)"=   c(0, 0, 0, 0, 0, -1, 1))
# Multiple comparisons and simultaneous confidence intervals
comps2 <- glht(modell1, linfct = mcp(Treatment = contrasts2),
  alternative="two.sided")
SCI2<-confint(comps2)
# Create the plot in Fig. 2
windows(8,5); par(mar=c(5, 12, 4, 1) + 0.1)
plot(SCI2, main="", xlab="Difference in fruit set
(number of fruits per 100 flower clusters)")
```

The following R code reproduces the calculation of simultaneous confidence intervals plotted in Figure 3:

```
# Reorder the factor levels in the data set and fit the linear model
ExKCotton$KFertilizer <- factor(ExKCotton$KFertilizer,
levels=c("BC0BN0", "BC68BN0", "BC136BN0", "BC204BN0", "BC0BN34",
"BC0BN68", "BC0BN102", "BC68BN34", "BC68BN68", "BC136BN34", "BC136BN68"))
model2 <- lm(Yield~KFertilizer, data=ExKCotton)
# Define the contrasts in Table 7
contrasts3 <- rbind(
"1. Bc68Bn0 - Bc0Bn0" = c(-1,1,0,0,0,0,0,0,0,0,0,0),
"2. Bc136Bn0 - Bc0Bn0" = c(-1,0,1,0,0,0,0,0,0,0,0,0),
"3. Bc204Bn0 - Bc0Bn0" = c(-1,0,0,1,0,0,0,0,0,0,0,0),
"4. Bc0Bn34 - Bc0Bn0" = c(-1,0,0,0,1,0,0,0,0,0,0,0),
"5. Bc0Bn68 - Bc0Bn0" = c(-1,0,0,0,0,1,0,0,0,0,0,0),
"6. Bc0Bn102 - Bc0Bn0" = c(-1,0,0,0,0,0,1,0,0,0,0,0),
"7. Bc68Bn34 - Bc0Bn0" = c(-1,0,0,0,0,0,0,1,0,0,0,0),
"8. Bc68Bn68 - Bc0Bn0" = c(-1,0,0,0,0,0,0,0,1,0,0,0),
"9. Bc136Bn34 - Bc0Bn0" = c(-1,0,0,0,0,0,0,0,0,1,0,0),
"10. Bc136Bn68 - Bc0Bn0" = c(-1,0,0,0,0,0,0,0,0,0,1,0),
"11. Bc68Bn34 - Bc0Bn102" = c(0,0,0,0,0,0,-1,1,0,0,0,0),
"12. Bc68Bn68 - Bc136Bn0" = c(0,0,-1,0,0,0,0,0,0,1,0,0),
"13. Bc136Bn68 - Bc204Bn0" = c(0,0,0,-1,0,0,0,0,0,0,1,0),
"14. Bc0Bn68 - Bc68Bn0" = c(0,-1,0,0,0,1,0,0,0,0,0,0))
# Multiple comparisons and simultaneous confidence intervals
comps3 <- glht(model2, linfct = mcp(KFertilizer = contrasts3),
alternative="greater")
SCI3<-confint(comps3)
# Create the plot in Fig. 3
windows(8,5); par(mar=c(5, 12, 4, 1) + 0.1)
plot(SCI3, main="", xlab="Difference in cotton yield (kg/ha)")
```