Psychological Bulletin
1988, Vol. 104, No. 1, 145–149

0033-2909/88/$00.75

Improved Bonferroni-Type Multiple Testing Procedures

Burt S. Holland
Department of Statistics
Temple University

Margaret DiPonzio Copenhaver
Wyeth-Ayerst Laboratories
Philadelphia, Pennsylvania

The Bonferroni multiple comparisons procedure is customarily used when doing several simultaneous tests of significance in relatively nonstandard situations in which other methods do not apply. We review some new and improved competitors to the Bonferroni procedure, that although constraining generalized Type I error probability to be at most α, afford increased power in exchange for increased complexity in implementation. An improvement to the weighted form of the Bonferroni procedure is also presented. Several data sets are reanalyzed with the new methods.

It is widely recognized that when one wishes to simultaneously test each of several hypotheses at a common significance level α', the generalized Type I error probability α, (i.e., the probability of rejecting at least one of those hypotheses being tested that is in fact true), is typically much in excess of α'. Thus as noted by Myers (1979) and Winer (1971), it is usually desirable to ensure a preselected value for α by using a multiple comparisons testing procedure for conducting the multiple tests. As recent surveys will attest, there are a large number of multiple testing procedures available (Hochberg & Tamhane, 1987; Jaccard, Becker, & Wood, 1984; Miller, 1981).

Our concern in this article is with procedures flexible enough to deal with relatively nonstandard testing situations. Examples include the following:

1. One has M populations with associated parameter θ_i from population i, $i = 1, 2, \ldots, M$, and one wants to conduct tests of the form H_0: $\theta_i = \theta_{i'}$ $(i \neq i')$ on considerably fewer than all $M(M-1)/2$ possible pairs of such parameters. An important instance of this setup occurs where μ_{gt} is the mean of an observation from group g at the tth treatment level ($g = 1, 2, \ldots, G$; $t = 1, 2, \ldots, T$) and one wishes to test only each treatment within group ("simple effect") hypothesis of the form H_0: $\mu_{gt} = \mu_{gt'}$, $g = 1, 2, \ldots, G$; $1 \leq t < t' \leq T$; see Copenhaver and Holland (1987). Here there are $GT(GT-1)/2$ possible pairs of means one can test, but interest lies only in a $GT(T-1)/2$-sized subset of them.

2. An analysis of variance (ANOVA) table contains many F tests. Unless the experimental design is balanced and the error degrees of freedom is quite large, these F statistics are statistically dependent and percentage points (or algorithms for constructing them) for the required multivariate distributions are not readily available.

3. In an $r \times c$ contingency table one wishes to conduct tests of independence on some or all of the $rc(r-1)(c-1)/4$ possible 2×2 subtables.

In instances such as these, researchers typically invoke the Bonferroni procedure wherein if there are k hypotheses to be tested, each test should be conducted at significance level α/k, or more generally, hypothesis i should be tested at level α_i, $0 < \alpha_i < 1$, where the $\{\alpha_i\}$ are chosen to satisfy $\sum_{i=1}^{k} \alpha_i = \alpha$. For a discussion of this approach see Rosenthal and Rubin (1984).

Recently several authors have introduced improved versions of the Bonferroni procedure that should be considered for use in most situations in which Bonferroni has heretofore been suitable. These improvements provide varying nonnegligible increases in the power of the tests conducted. In exchange for the higher power, the newer procedures are more complex to invoke than Bonferroni, but assuming computer implementation, this is a relatively minor issue. Each of these improvements assumes the availability of the p value (observed significance level) for each individual test undertaken, a reasonable assumption nowadays. However, because of their stagewise nature, the newer procedures, unlike Bonferroni, cannot be accompanied by analogous confidence intervals.

First in this article the improved procedures are presented, illustrated with numerical examples. We then review how one of these procedures can be extended to a weighted form, thus constituting an increased power improvement on the weighted Bonferroni procedure discussed by Rosenthal and Rubin (1984) and others.

We denote the k hypotheses $H_1, \ldots, H_k$ and assume these are a "minimal set"; that is, no hypothesis in the set is equivalent to the simultaneous occurrence of two or more other hypotheses in the set. Then α as defined earlier is the probability of rejecting at least one H_i that is in fact true. For each $i = 1, 2, \ldots, k$ let X_i be the test statistic and let p_i be the p value of the test of H_i.

Holm's Procedure

Order the k p values from smallest to largest and denote the ordered $\{p_i\}$ by $p_{(1)} \leq \cdots \leq p_{(k)}$. Tied p_i can be ordered arbitrarily. Let $H_{(1)}, \ldots, H_{(k)}$ be the corresponding hypotheses. Suppose i^* is the smallest integer from 1 to k such that

We are grateful to the referees for several helpful comments.

The work of Burt S. Holland was supported in part by a Temple University research and study leave.

Correspondence concerning this article should be addressed to Burt S. Holland, who is now at Merck Sharp & Dohme Research Laboratories, West Point, Pennsylvania 19486.

Table 1
P Values for Case-Control Study

i	$p_{(i)}$	$\alpha/(k-i+1)$
1	.0001	.0042
2	.0002	.0045
3	.0012	.0050
4	.0022	.0056
5	.0033	.0063
6	.0084	.0071
7	.016	.0083
8	.0273	.0100
9	.033	.0125
10	.039	.0167
11	.047	.0250
12	.048	.0500

$$p_{(i^*)} > \alpha/(k - i^* + 1). \quad (1)$$

Then Holm's (1979) procedure rejects $H_{(1)}, \ldots, H_{(i^*-1)}$ and retains $H_{(i^*)}, \ldots, H_{(k)}$. If Equation 1 is true for no integer i^* then all k hypotheses are rejected. Holm proves that this procedure guarantees that there is at most an α chance of rejecting at least one of the true hypotheses.

Since the usual Bonferroni procedure may be described as in the previous paragraph except for replacement of $\alpha/(k - i^* + 1)$ in Equation 1 with α/k, a number smaller for $i^* > 1$, any hypothesis rejected by Bonferroni is also rejected by Holm. The increased power arises from the fact that Holm may reject some hypotheses not rejected by Bonferroni. In practice this is a distinct possibility since $\alpha/(k - i + 1)$ is much larger than α/k for large values of i and if $p_{(i)}$ lies between $\alpha/(k - i + 1)$ and α/k, and $i < i^*$, then $H_{(i)}$ is retained by Bonferroni but rejected by Holm.

Application to a Case-Control Study With Both t *and Chi-Square Tests*

Solkoff, Gray, and Keill (1986) studied posttraumatic stress disorder among Vietnam veterans. In their Table 2 they present the results of tests of significance for a family of 12 case-control comparisons dealing with combat experience. Eight of the individual tests were chi-squares and the other four were one-tailed *t*s.

The authors reported the individual p values for the $k = 12$ individual tests; these are reproduced in Table 1, listed in order of increasing p value. No attempt was made in their analysis to control the generalized Type I error probability α.

Suppose we reconsider these data controlling α at .05. Referring to Table 1, see that i^* in Equation 1 equals 6 (.0033 < .0063 but .0084 > .0071). Hence, testing at generalized Type I error probability α, one rejects $H_{(1)}, H_{(2)}, \ldots, H_{(5)}$ and retains $H_{(6)}, H_{(7)}, \ldots, H_{(12)}$.

Shaffer's Procedures

Shaffer (1986) offered two enhancements to Holm's procedure. The second is more difficult to implement than the first but affords somewhat larger power.

In Shaffer's first procedure, hereafter denoted S1, t_i is defined to be the maximum number of possibly true null hypotheses conditional on the assumption that at least $i - 1$ of the k hypotheses are false; examples of the construction of the $\{t_i\}$ are given below. Let i^* be the smallest integer from 1 to k such that

$$p_{(i^*)} > \alpha/t_{i^*}. \quad (2)$$

Then S1 rejects $H_{(1)}, \ldots, H_{(i^*-1)}$ and retains $H_{(i^*)}, \ldots, H_{(k)}$. If Equation 2 is true for no integer i^* then all k hypotheses are rejected.

If $i - 1$ hypotheses are false then at most all of the remaining $k - i + 1$ hypotheses are true. But if the hypotheses are logically interrelated (i.e., the truth of some of the hypotheses necessarily implies the truth of some others), then $t_i < k - i + 1$ for some values of i. For example, suppose we have four populations with corresponding parameters $\theta_1, \theta_2, \theta_3, \theta_4$, and we wish to test all six hypotheses of the form $\theta_j = \theta_{j'}$ ($j \neq j'$). Further suppose that in actuality at least one of the six hypotheses is false. Then necessarily at most three of the remaining five hypotheses can be true; hence $t_2 = 3$.

Since $t_i \leq k - i + 1$, the right side of Equation 2 may exceed the right side of Equation 1. Thus S1 rejects all hypotheses rejected by Holm and possibly some additional ones; in other words, the Shaffer procedure is at least as powerful as Holm's.

Clearly, generation of the $\{t_i\}$ is required for implementation of S1. The difficulty of this task, which is not required by Holm's procedure, depends heavily on the specific testing situation. Universally applicable guidelines for constructing the $\{t_i\}$ cannot be given, but construction is illustrated below for a particular example.

The second procedure of Shaffer, which we denote S2, offers a small increase in power compared with S1 in exchange for a further increase in complexity. Here Shaffer redefines t_i to be the maximum number of possibly true null hypotheses conditional on the assumption that the specific hypotheses $H_{(1)}, \ldots, H_{(i-1)}$ are false. Then the test proceeds as in Equation 2. After this redefinition t_i is possibly smaller than it was in S1. Hence the right side of Equation 2 in S2 may exceed that in S1 so S2 is at least as powerful as S1. Shaffer proves that both versions of her procedure ensure the generalized Type I error probability to be at most α.

We illustrate choice of the $\{t_i\}$ under S1 and S2 with a small example. Assume there are six populations with associated parameters $\theta_1, \ldots, \theta_6$, and suppose we wish to test only the $k = 6$ hypotheses:

$$H_1\colon \theta_1 = \theta_2, \quad H_2\colon \theta_1 = \theta_3, \quad H_3\colon \theta_2 = \theta_3,$$
$$H_4\colon \theta_4 = \theta_5, \quad H_5\colon \theta_4 = \theta_6, \quad H_6\colon \theta_5 = \theta_6. \quad (3)$$

Obviously, for either S1 or S2, $t_1 = 6$, and if one hypothesis is false then at most four are true; hence $t_2 = 4$. However, the value of t_3 may differ between S1 and S2 depending on the configuration of $H_{(1)}$ and $H_{(2)}$. For instance, if $H_{(1)}$ is one of H_1, H_2, or H_3, and $H_{(2)}$ is one of H_4, H_5, or H_6, then t_3 for S2 is 2; if $H_{(1)}$ and $H_{(2)}$ are both among H_1, H_2, and H_3, then t_3 for S2 is 4. It is easily seen that t_3 for S1 is 4, the larger of the two possible choices in this example under S2. Observe that generation of $\{t_i\}$ under S2 is data dependent—it requires tracking the specific hypotheses in the order in which they are tested, which in turn depends on the ordering of the p values.

The Procedures of Holland and Copenhaver

Holland and Copenhaver (1987) have supplied an enhancement, abbreviated here as HC, adaptable to each of the foregoing procedures. Its applicability is slightly less general than that of Holm's or either Shaffer procedure, but since its use entails no new details for the analysis, it should be used whenever its required assumption of positive orthant dependence of the test statistics is met.

Suppose, as is usual, that each test statistic X_i can be written such that large values indicate rejection of H_i. The positive orthant dependence condition on the underlying multivariate distribution states

$$pr(X_1 \leq x_1, \ldots, X_k \leq x_k) \geq \prod_{i=1}^{k} pr(X_i \leq x_i) \quad \text{for all } x_1, \ldots, x_k.$$

This condition applies in most practical testing situations involving the normal: Student's t, chi-square, F, and other common distributions, including the three "nonstandard" examples offered earlier. However, it may fail to hold when some of the individual tests are of the one-tailed variety involving the Student's t or normal distributions. Further discussion is provided by Holland and Copenhaver (1987).

Define $C(x) = 1 - (1 - \alpha)^{1/x}$ and note that $C(x) \geq \alpha/x$ for any $x \geq 1$, with equality only for $x = 1$. The HC enhancement to Holm entails replacement of the right side of Equation 1 with the larger number $C(k - i^* + 1)$, and the HC enhancement to S1 and S2 consists of replacing the right side of Equation 2 with the larger number $C(t_{i^*})$. It is seen that these replacements raise the possibility that HC may reject some hypotheses that the unrefined procedures do not; hence HC may provide increased power. Holland and Copenhaver (1987) proved that their enhancement, when applicable, guarantees that there is at most an α chance of generalized Type I error.

The HC correction is analogous to the widely used Sidak (1967) modification of the Bonferroni procedure for simultaneously testing equality of means with multivariate t statistics; see Miller (1981). Although the potential for gain in power as a result of using HC is slight since $C(x)$ is only slightly greater than α/x for most α and $x > 1$, its use seems worthwhile because calculation of $C(x)$ is effortless.

Application to F *Tests in a Four-Way Analysis of Variance*

Marsh (1986) presented three $2 \times 2 \times 3 \times 4$ ANOVAS. The response variables were measures of students' perceptions of the cause of their academic successes and failures; the factors were academic content (C), outcome (O), perceived cause (P), and level of mathematics achievement (L). For each analysis, 15 F tests were run for the four main effects and 11 interaction effects. Suppose we reevaluate the results from Marsh's Table 1 (Study 3), setting generalized Type I error probability at .05.

Eight of the tests had p values of less than .001, so their null

Table 2
HC-Enhanced S2 Analysis of Behavioral Adjustment Data With α = .10

i	t_i	$p_{(i)}$	$1-(1-\alpha)^{1/t_i}$	Psychiatric category	Method of therapy
1	10	.0006	.0105	4	1 vs. 2
2	10	.0049	.0105	2	2 vs. 3
3	8	.0151	.0131	2	1 vs. 3

Note. HC = Holland & Copenhaver (1987); S2 = Shaffer's second procedure (1986).

hypotheses would be rejected according to the least powerful of the procedures we have discussed, Bonferroni, which requires $p_i > .05/15$ for acceptance. Six of the tests had p values of at least .24; these correspond to hypotheses that would unquestionably be retained by all procedures. The remaining test, for $C \times O \times P$ interaction, had $p_{(9)} = .005$. Hypothesis $H_{(9)}$ is retained by Bonferroni because $p_{(9)} > .05/15$. However, this hypothesis is rejected by the HC modification of Holm's procedure since $p_{(9)} < 1 - (1 - .05)^{1/(15-9+1)} \doteq .0073$, thereby illustrating the increased power of HC over Bonferroni when applied to this data set. $H_{(9)}$ is also rejected by Holm, S1, and S2; for each of these procedures the critical value is $.05/7 \doteq .0071$.

Application to Simple Effects Tests in a Two-Way Analysis of Variance

Denenberg (1976) presented a data set in which patients from each of four psychiatric categories (P) were randomly assigned to one of three methods of therapy (M). The response variable was a composite score representing "degree of adjustment of behavior."

The ANOVA table indicated a significant $P \times M$ interaction ($p = .026$). We will use the HC modification of S2 to compare the therapy methods within each level of P in a pairwise manner for a total of 12 comparisons. We fixed the generalized Type I error probability at .10.

In the first stage of testing t_1 typically equals 12. However, as discussed by Shaffer (1986), since a preliminary interaction test was performed and was significant, one can proceed as if at least one cell mean differs from the remaining two within one of the levels of P. The value of t_1 can thus be reduced from 12 to 10 without changing the generalized error rate.

The results for the first three stages of testing appear in Table 2. At Stage 3, because $H_{(1)}$ and $H_{(2)}$ correspond to hypotheses within different levels of P, there remains at most one hypothesis in Levels 4 and 2 that can be true and at most three hypotheses in the remaining two levels that can be true. Thus $t_3 = 8$. At this point testing stops since .0151 > .0131. We conclude that if one sets generalized Type I error probability at .10 for the 12 simple effects comparisons, the only significant differences are M1 versus M2 at P4 and M2 versus M3 at P2.

Holm's Improved Weighted Bonferroni Procedure

Rosenthal and Rubin (1984) describe a Bonferroni-like procedure, (hereafter abbreviated RR) in which a relatively larger amount of overall α may be allotted to those hypotheses deemed relatively "more important" in order to increase the power of the test of such hypotheses. If $w_1, w_2, \ldots, w_k$ are such positive weights for the k hypotheses, with larger w_i for "important" hypotheses and smaller w_i for "unimportant" hypotheses, it is advocated that the ith p value be compared with $\alpha w_i / \sum_1^k w_i$ instead of with α/k as in the usual Bonferroni procedure.

This use of weights in multiple significance testing is controversial—see for instance the debate between Ryan (1985) and Rosenthal and Rubin (1985). What seems clear to all is that the weights chosen must be clearly justified prior to analyzing the data; researchers are cautioned not to analyze or reanalyze their data in order to find a set of weights that will produce results supporting conclusions pleasing to themselves. One way to discourage indiscriminate use of weighted procedures would be to routinely require such presentations to be accompanied by a sensitivity analysis in which indications are given of what types of choices of $\{w_1, w_2, \ldots, w_k\}$ will give rise to each possible overall conclusion.

Holm's (1979) weighted sequentially rejective procedure, hereafter abbreviated WH, provides an increased power improvement on RR analogous to the improvement of his unweighted procedure, discussed previously, over the ordinary Bonferroni method. For each hypothesis H_i, $i = 1, 2, \ldots, k$, Holm defines the p-value-to-weight ratio $r_i = p_i/w_i$. Let $r_{(1)} \le r_{(2)} \le \cdots \le r_{(k)}$ be the ordered r_is, let $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$ be the corresponding hypotheses, and let $w_{(1)}, w_{(2)}, \ldots, w_{(k)}$ be the corresponding weights. Further let $p_{(1)}, p_{(2)}, \ldots, p_{(k)}$ be the corresponding p values, but note that these are not the ordered p values as in earlier sections of this paper. That is, $w_{(i)}$ and $p_{(i)}$ are defined through $r_{(i)}$, and $H_{(i)}$ is the hypothesis whose p value is in the numerator of $r_{(i)}$. Then Holm's weighted procedure is as follows. Let i^* be the smallest integer such that

$$r_{(i^*)} > \alpha \Big/ \sum_{m=i^*}^{k} w_{(m)}. \qquad (4)$$

Then reject $H_{(1)}, \ldots, H_{(i^*-1)}$ and retain $H_{(i^*)}, \ldots, H_{(k)}$. If no integer i^* satisfies Equation 4 then all k hypotheses should be rejected.

To see that WH has higher power than RR, suppose RR rejects $H_{(i)}$. Then

$$p_{(i)} < \frac{\alpha w_{(i)}}{\sum_{m=1}^{k} w_{(m)}}.$$

This immediately leads to

$$r_{(i)} < \frac{\alpha}{\sum_{m=1}^{k} w_{(m)}} < \frac{\alpha}{\sum_{m=i}^{k} w_{(m)}},$$

and thus WH also rejects $H_{(i)}$. Indeed, if $p_{(i)}$ falls in the interval

$$\left(\frac{\alpha w_{(i)}}{\sum_{m=1}^{k} w_{(m)}}, \frac{\alpha w_{(i)}}{\sum_{m=i}^{k} w_{(m)}} \right) \qquad (5)$$

and if $i < i^*$, then $H_{(i)}$ is retained by RR but rejected by WH. The Interval 5 becomes increasingly large as (i) increases. Hence WH may well detect true differences for hypotheses with relatively larger p-value-to-weight ratio; true differences that are not detected by RR.

Holm (1979) also proved that WH constrains the generalized Type I error probability to be at most α.

Table 3
Calculations for Holm's Weighted Procedure Analysis of Pupil Intellectual Performance Data

i	Effect	$w_{(i)}$	$p_{(i)}$	$p_{(i)}/w_{(i)}$	$\alpha / \sum_{m=i}^{k} w_{(m)}$
1	EA	3	.0007	.00023	.00217
2	EP	3	.001	.00033	.00250
3	D	1	.0007	.0007	.00294
4	DA	1	.0007	.0007	.00313
5	EDA	1	.001	.001	.00333
6	PD	1	.001	.001	.00357
7	E	3	.031	.01033	.00385
8	ED	3	.042	.014	.00500
9	EPD	1	.030	.030	.00714
10	A	1	.030	.030	.00833
11	PDA	1	.030	.030	.01000
12	EPA	1	.031	.031	.01250
13	EPDA	1	.031	.031	.01667
14	P	1	.042	.042	.02500
15	PA	1	.042	.042	.05000

Note. E, P, D, and A = the four factors.

Reanalysis of Pupil Intellectual Performance Data Using Holm's Weighted Procedure

This data set, from a 2^4 factorial experiment, was used by Rosenthal and Rubin (1984) to illustrate what we have termed their RR procedure. We use it here to demonstrate the calculations for the WH procedure.

The four factors are abbreviated E, P, D, and A. As in Rosenthal and Rubin, we assume a generalized Type I error probability of .05 and that the effects E, EP, ED, and EA each receive a weight of 3 while the other 11 effects carry a weight of 1. The p values from Rosenthal and Rubin are listed in the fourth column of Table 3 in order of increasing p-value-to-weight ratio displayed in the fifth column.

The calculations in this table indicate that the test statistics for the first six effects (EA, EP, D, DA, EDA, and PD) are statistically significant at generalized Type I error probability $\alpha = .05$ while the other nine test statistics are not (.001 < .00357 but .01033 > .00385). This happens to be the same conclusion reached by RR in this example, but in general there is an appreciable chance that WH will reject some hypotheses that RR would retain.

Discussion

We recommend that a procedure selected from among those presented herein be used in preference to the unweighted or weighted Bonferroni procedure in simultaneous testing situations where Bonferroni would otherwise be the method of choice, provided that p values for each test are available and confidence intervals are not required. If the positive orthant dependence condition can be verified, the HC enhancement should be implemented.

When searching the literature for numerical examples in preparation of this article, we noted numerous instances, particularly in regard to F tests, of multiple testing in which authors were oblivious to the erroneous rejection of hypotheses occasioned by lack of control of the inherent multiplicity of tests. It is hoped that researchers will give increased attention to this problem.

We were also struck by the current common use of the Newman-Keuls multiple comparisons procedure, which to a large extent has fallen out of favor. The Newman-Keuls procedure controls at level α the probability of rejecting at least one null hypothesis if all null hypotheses are actually true. But if some hypotheses are true and others are false, then the probability that the Newman-Keuls procedure rejects at least one true hypothesis may be considerably greater than α.

References

Copenhaver, M. D., & Holland, B. (1987, August). *Multiple comparisons of simple effects in a balanced two-way layout.* Paper presented at Joint Statistical Meetings, San Francisco, CA.

Denenberg, V. H. (1976). *Statistics and experimental design for behavioral and biological researchers.* New York: Wiley.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparisons procedures.* New York: Wiley.

Holland, B., & Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics, 43,* 417–423.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65–70.

Jaccard, J., Becker, M. A., & Wood, G. (1984). Pairwise multiple comparison procedures: A review. *Psychological Bulletin, 96,* 589–596.

Marsh, H. W. (1986). Self-serving effect (bias?) in academic attribution: Its relation to academic achievement and self-concept. *Journal of Educational Psychology, 78,* 190–200.

Miller, R. G., Jr. (1981). *Simultaneous statistical inference* (2nd ed.). New York: Springer-Verlag.

Myers, J. L. (1979). *Fundamentals of experimental design* (3rd ed.). Boston: Allyn and Bacon.

Rosenthal, R., & Rubin, D. B. (1984). Multiple contrasts and ordered Bonferroni procedures. *Journal of Educational Psychology, 76,* 1028–1034.

Rosenthal, R., & Rubin, D. B. (1985). Statistical analysis: Summarizing evidence versus establishing facts. *Psychological Bulletin, 97,* 527–529.

Ryan, T. A. (1985). Ensemble-adjusted p values: How are they to be weighted? *Psychological Bulletin, 97,* 521–526.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association, 81,* 826–831.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62,* 626–633.

Solkoff, N., Gray, P., & Keill, S. (1986). Which Vietnam veterans develop posttraumatic stress disorder? *Journal of Clinical Psychology, 42,* 687–698.

Winer, B. J. (1971). *Statistical principles in experimental design.* New York: McGraw-Hill.

Received May 26, 1987
Revision received September 4, 1987
Accepted September 14, 1987 ■