A Study of the Powers of Several Methods of Multiple Comparisons

ISRAEL EINOT and K. R. GABRIEL*

Powers of multiple comparisons procedures are studied for fixed maximal experimentwise levels. Analytical considerations show Tukey-Scheffé methods to have least power, Duncan's to be intermediate, Ryan's most powerful. (Newman-Keuls tests could preserve experimentwise levels only if modified radically and impractically.) Extensive Monte-Carlo trials show these power differences to be small, especially for range statistics. We therefore generally recommend the Tukey technique for its elegant simplicity and existent confidence bounds—its power is little below that of any other method. Simulation was for 3, 4 and 5 treatments: the conclusions might need modification for more treatments.

1. THEORY

1.1 Introductory Remarks[1]

The purpose of this article is to compare some current multiple comparisons procedures (MCPS) in terms of power, when the probability of making no Type I error at all is kept above a common lower bound. The discussion is restricted to the one-way normal ANOVA set-up and deals with well-known MCPS which provide decisions of rejection or retention of hypotheses of homogeneity for all subsets of treatments.

All known MCPS ensure the *coherence* [10, p. 229] of their decisions in the sense that they cannot consider a set of treatments as possibly homogeneous if any subset of it has been declared heterogeneous. Most MCPS ensure this coherence simply by proceeding in a stepwise manner from sets to their subsets and automatically declaring homogeneity for each subset of a homogeneous set. (Section 1.6.) Simultaneous test procedures (STPS) are structured in such a manner that the stepwise sequence of testing is not necessary.

Most MCPS cannot avoid *dissonances* [10, p. 231] such as declaring a set heterogeneous without doing so for any one of its proper subsets. Because of the possibility of such errors one should not speak of "accepting" homogeneity hypotheses, but merely of retaining them. This is a general feature of all statistical tests, not only of MCPS.

A reasonable yardstick for the likelihood of false declarations of heterogeneity, i.e., of making Type I errors, is the probability of making any such errors at all, on all subsets. When all true treatments are equal, this is called the *experimentwise level* of the MCP and has found much acceptance as a stringent but clear-cut criterion for the proneness of MCPS to Type I errors. We denote it by α. Because of the MCPS' coherence (whenever there is any rejection there must be one on the total set), this level is readily seen to be equal to the Type I error probability of the decision on the total set of all means.

We take exception to the universal relevance of the experimentwise level as a criterion for judging the false rejection probability of an MCP. We do so because MCPS may have false rejection probabilities which are well above their α's [29; 24, p. 30]. Of the MCPS discussed here, only those based on the Newman-Keuls allocation of levels suffer from this flaw (see Section 1.7). We have chosen to formulate this study of MCPS in terms of a constant experimentwise level and, therefore, must consider the Newman-Keuls procedures invalid unless modified appropriately (see Section 1.8).

We have preferred this course to formulating our comparisons in terms of an unfamiliar measure of maximal probability of false rejection.

Probabilities of false rejection, or *levels*, can also be formulated for different subsets (see Section 1.3). Unfortunately, different authors have presented their MCPS in terms of levels for different sizes of subsets. Thus, for example, Scheffé [26] has stressed the experimentwise level—i.e., the level for the test of all means—whereas critical values for Duncan's method are usually tabulated in terms of the pairwise comparison error rate [14, 15]. Clearly, the different methods are not comparable when each one is studied with a probability of say, five percent of the kind of Type I errors discussed in its original presentation. Obviously, the Scheffé method at the five-percent experimentwise level will yield many fewer significant results than Duncan's method at the five-percent level for each pairwise comparison.

Strangely enough, comparative studies of multiple comparisons methods have ignored this point and have compared all methods by fixing each at the Type I error probability at which it was published. Whether such studies compare only significance levels [3] or levels and powers [2, 4, 23], they are misleading in that most of

* Israel Einot is assistant and K.R. Gabriel is professor, both at Department of Statistics, Hebrew University, Jerusalem, Israel. The research was supported in part by grant NCHS-IS-1 of the U.S. National Center for Health Statistics. This study contains the results of the first author's M.Sc. thesis, Hebrew University, directed by the second author.

[1] We gratefully acknowledge the many discussions we have had with Dr. Joseph Putter, which helped us to a clearer understanding of the problems of multiple comparisons.

their conclusions are simple consequences of the choices of Type I error probabilites, rather than of the techniques used. No Monte Carlo study was needed to realize the alleged "inferiority of the Scheffé, Tukey and Newman-Keuls procedures for detecting real differences" [4, p. 73]. This misconstruction arose simply because these MCPS were set at experimentwise $\alpha = 0.05$ and compared with several other procedures using a five-percent probability of false rejection in pairwise comparisons. Similar conclusions could be inferred from Harter's approximate [13] and exact [14] power calculations, again setting different types of levels at five percent.

The import of this discussion and Monte Carlo results (Section 2) is to allow comparison of MCPS, all of which have the same experimentwise level α. The MCPS are therefore compared on the statistics used, on the allocation of the experimentwise level to subsets of differing sizes and on the procedures for actually arriving at the decisions on all subsets.

1.2 A General Formulation of MCPs

To clarify ideas and concepts, we begin with a formulation of MCPS sufficiently general to include all those discussed here. The following data are assumed to be available: independent normally distributed means $\bar{y}_1, \cdots, \bar{y}_k$ based on samples from k populations. These means have variances $\sigma^2/n_1, \cdots, \sigma^2/n_k$, for known sample sizes $n_1, \cdots, n_k$, respectively, and unknown σ^2. An estimate s^2 of σ^2 is available such that $n_e s^2/\sigma^2$ is chi-square with n_e degrees of freedom and is independent of the means. An MCP provides tests of the equality of all the true means in each set P of $p(= 2, \cdots, k)$ of the k treatments. The test of homogeneity of set P—hypothesis ω_P—uses a statistic

$$T_P = T_p(\bar{y}_i, i \in P; s^2) \tag{1.1}$$

based symmetrically on the p means of set P and on the variance estimate s^2. (Symmetry here means that relabeling of the means $\bar{y}_i, i \in P$ will not affect the value of the statistic.) Statistic T_P is required to be such that its distribution under ω_P depends on set P only through the number p of treatments in P.

All the MCPS under discussion compare the statistics T_P for all sets P of size $p(= 2, \cdots, k)$ with the same critical value ζ_p. An MCP thus requires the definition of $k - 1$ critical values $\zeta_2, \cdots, \zeta_k$ for the comparison of k means.

The hypothesis of homogeneity of set P is rejected if $T_P > \zeta_p$ as well as $T_R > \zeta_r$ for all sets R of $r(= p + 1, \cdots, k)$ treatments which contain P. In other words, ω_P is rejected if and only if

$$T_R > \zeta_r \forall R (P \subseteq R \subseteq K) \ , \tag{1.2}$$

where K is the total set of all k treatments. Conversely, ω_P is retained if $T_R \leq \zeta_r$ for at least one set R containing P, including P itself and the total set K.

The actual checking of the events $T_R > \zeta_r$ for all R containing the set P is carried out in a stepwise manner (see Section 1.6). This cross checking is essential to the coherence of the decisions on different subsets.

1.3 Levels and Powers

Nominal significance level γ_p is defined as

$$\gamma_p = P_{\omega_p}(T_P > \zeta_p) \tag{1.3}$$

for any set P of size $p(= 2, \cdots, k)$. This is the probability that the statistic T_P for a homogeneous set P of size p will exceed critical value ζ_p. Since the distribution of T_P has been noted to be the same for all such sets of given size p, the probability in (1.3) also depends only on p. This is indicated by the notation γ_p.

Nominal level γ_p is seen to be, for any given p, a monotonically decreasing function of critical value ζ_p. Except for points of discontinuity, ζ_p uniquely determines γ_p and vice versa. An MCP with given statistics T of (1.1) may, therefore, be defined equally well by the set of critical values $\zeta_2, \cdots, \zeta_k$ or by the corresponding set of nominal levels $\gamma_2, \cdots, \gamma_k$.

The levels γ_p are called nominal because in MCPS hypothesis ω_P is not necessarily rejected when event $T_P > \zeta_p$ of (1.3) occurs (except with STPS; see (1.14)). As noted previously, ω_P will be retained despite $T_P > \zeta_p$ if, for some R containing P, $T_R \leq \zeta_r$. Hence, the true levels usually fall short of the nominal levels γ_p. The exception is for $p = k$, where $T_K > \zeta_k$ is clearly the sufficient condition for rejecting ω_K. Hence, γ_k also is the true level of the k treatments comparison and is equal to the experimentwise level, that is

$$\gamma_k = \alpha \ . \tag{1.4}$$

For any subset P the *true level*—that is, the probability of rejecting homogeneity hypothesis ω_P if true—clearly depends in part on the configuration of the true means of treatments outside P and of their relations to the common mean of the treatments in P (again, except for STPS). It is evident that the true level will be maximal if the means outside P greatly differ from those in P, as this makes it virtually certain that $T_R > \zeta_r$ whenever $R \supset P$ and thus makes the decision ω_P depend exclusively on whether $T_P > \zeta_p$. The true level is thus seen to be at most equal to the nominal level. (They are equal for STPS.)

Just as an MCP has a collection of levels, one for each subset, so it has a corresponding collection of powers against each alternative configuration of the true treatment means. As for the true levels, so also the power on any set P depends not only on the differences among the true means of the treatments in P but also on those of treatments outside P. Writing Ω for the configuration of all k true treatment means, one may write the power as

$$\mathcal{P}_\Omega(P) = P_\Omega(\bigcap_{P \subseteq R \subseteq K} (T_R > \zeta_r)) \ . \tag{1.5}$$

It is evident from (1.5) that these powers are inversely dependent on the set of critical values $\zeta_2, \cdots, \zeta_k$, just as the nominal levels γ_p (1.3). In fact, this makes the powers depend directly on the set of nominal levels $\gamma_2, \cdots, \gamma_k$.

1.4 Statistics and Allocation of Levels

Relations (1.3) and (1.5) between critical values ζ, nominal levels γ and powers β depend on the statistics T, that is, on the $k-1$ functions $T_p (p = 2, \cdots, k)$ of the p means and the variance estimate. The probabilities in (1.3) and (1.5) assume different values when different statistics are used. MCPs may differ because they use different nominal levels with the same statistics, or because they use different statistics at the same nominal levels, or because of both. Comparative studies of different methods can be meaningful only if it is quite clear how the different MCPs differ. It is therefore necessary to turn to an enumeration of the different statistics considered in this study and then to the different ways of allocating nominal levels.

Two statistics are most commonly used in multiple comparisons when the usual normal analysis of variance conditions can be assumed to hold. The first is the *Studentized range*

$$T_P^{(1)} = \max_{i,e \in P} \{(\bar{y}_i - \bar{y}_e)[\min (n_i, n_e)]^{\frac{1}{2}}\}/s \ . \quad (1.6)$$

For equal sample sizes

$$n_1 = n_2 = \cdots = n_k = n \ , \quad \text{say} \ , \quad (1.7)$$

this becomes the well-known

$$T_P^{(1)} = (\max_{i \in P} \bar{y}_i - \min_{e \in P} \bar{y}_e)\sqrt{n}/s \ , \quad (1.8)$$

which is often denoted by $q(p, n_e)$. The more general form (1.6) has been brought to our attention by R.G. Miller and has since been published [27]. The second statistic is the *Studentized sum of squares* or *augmented F ratio*

$$T_P^{(2)} = (\sum_{i \in P} n_i \bar{y}_i^2 - (\sum_{i \in P} n_i \bar{y}_i)^2 / \sum_{i \in P} n_i)/s^2 \quad (1.9)$$

which is equal to $(p-1)$ times the $F_{(p-1, n_e)}$ ratio.

It is well known that under (1.7)

$$T_P^{(2)} = 2\{T_P^{(1)}\}^2 \quad \text{when} \quad p = 2 \ , \quad (1.10)$$

but this does not mean that the decisions on parts are the same in range MCPs as in sum of squares MCPs—for these decisions depend also on what has been decided on larger sets, where $T_P^{(2)}$ is not a function of $T_P^{(1)}$.

It is well known that under ω_P the distribution of $T_P^{(1)}$ and of $T_P^{(2)}$ each depend only on p and n_e and not on the actual homogeneous set P. This is as required in Section 1.2. Statistic $T_P^{(1)}$, under (1.7), has the Studentized range distribution of p normal and n_e error degrees of freedom. If (1.7) does not hold, the corresponding augmented Studentized range distribution produces very slightly conservative tests [27]. For $T_P^{(2)}$, one requires $p-1$ times the percentage points of the F distribution with $p-1$ and n_e degrees of freedom.

The following allocations of nominal levels of significance will be compared in the present study assuming a given experimentwise level α. The Newman [21] and Keuls [17] allocation is

$$\gamma_p^{NK} = \alpha (p = 2, \cdots, k) \quad (1.11)$$

(superscripts will be used to indicate a particular method); Duncan [7, 8] has argued for "protection levels" which become

$$\gamma_p^{D} = 1 - (1-\alpha)^{(p-1)/(k-1)} \quad (p = 2, \cdots, k) \quad (1.12)$$

when adjusted so as to ensure experimentwise level α; Ryan's idea of "adjusted significance levels" [25][2] can be used in the form

$$\gamma_p^{R} = 1 - (1-\alpha)^{p/k} \ . \quad (1.13)$$

Tukey [28] and Scheffé [26] have advocated

$$\gamma_p^{\text{STP}} = P_{\omega_P}(T_p > \zeta) \quad (p = 2, \cdots, k) \quad (1.14)$$

where $\zeta = \zeta_k$ and is defined by

$$\alpha = P_{\omega_K}(T_K > \zeta) \ . \quad (1.15)$$

The last allocation is based on use of a unique critical value ζ for all statistics T_P and the resulting methods may be referred to as STPs [10, 11].

The unique critical value ζ for an STP can be evaluated from (1.15) by use of the distribution of T_K under ω_K [1, 9]. For any other method M, which uses the same statistics T, ζ_k^M is evaluated as in (1.15) but the remaining critical values $\zeta_p^M (p = 2, \cdots, k-1)$ are, by (1.3), taken as upper $100\gamma_p^M$ percentage points of the distribution of T_P under ω_P. It has been shown (see [9, Sec. 9]) that

$$\zeta_p^M < \zeta_r^M \quad \text{if} \quad p < r \ , \quad (1.16)$$

so that the critical values increase with the size of set P to be tested.

1.5 Comparison of Levels, Critical Values and Powers

It is readily confirmed that

$$\alpha > 1 - (1-\alpha)^{p/k} > 1 - (1-\alpha)^{(p-1)/(k-1)} \quad (p = 2, \cdots, k-1) \quad (1.17)$$

with equalities for $p = k$. It then follows that

$$\gamma_p^{NK} > \gamma_p^{R} > \gamma_p^{D} \quad (p = 2, \cdots, k-1) \ , \quad (1.18)$$

and again for $p = k$ these are obviously equal to α.

It also follows that

$$\zeta_p^{NK} < \zeta_p^{R} < \zeta_p^{D} < \zeta_k \quad (p = 2, \cdots, k-1) \ , \quad (1.19)$$

and for $p = k$ the critical value for all methods is equal to ζ.

As noted in Section 1.3, power comparisons are related directly to comparisons of nominal levels and inversely to comparisons of critical values. It therefore follows that, for any one statistic T and any configuration Ω of the k

[2] Ryan himself chose the very slightly more conservative

$$\gamma_p = \alpha p/k \quad (1.13')$$

as it is easier for computation.

populations,

$$\beta_\Omega{}^{\text{STP}}(P) \geq \beta_\Omega{}^{R}(P) \geq \beta_\Omega{}^{D}(P) \geq \beta_\Omega{}^{\text{STP}}(P) \ , \quad \forall P \ . \quad (1.20)$$

In other words, whatever the configuration of populations and whatever statistic is used, the power of the MCP for any one subset P is always largest with the Newman-Keuls allocation. Next largest is the power with Ryan's allocation followed by that with Duncan's. STPs always have least power. In particular, such is the order for the true levels of significance.

These conclusions result from comparing all allocations at the same experimentwise level. Earlier studies have set the Duncan method at $\gamma_2 = 0.05$, the Newman-Keuls method at $\gamma_p = 0.05$ $(p = 2, \cdots, k)$ and the STP at $\gamma_k = 0.05$. Unsurprisingly, this has resulted in large levels and powers for Duncan's method, moderate ones for the Newman-Keuls method and low ones for the STP. The resulting Duncan, Newman-Keuls, STP order of powers was an obvious consequence of the differential setting of levels and had nothing to do with the intrinsic features of the different statistics and allocations. Harter's [13, 15] comparative conclusions and Balaam's [2], Petrinovich and Hardyck's [23] and Carmer and Swanson's [4] Monte Carlo studies all rediscover this ordering. Their conclusions and recommendations differ merely upon whether they are more concerned with increasing power or with protection against false rejection.

The different requirements of these various authors could have been attained with any one of the MCPs by choosing different levels according to each author's preferences. The more power-concerned analysts could have set the experimentwise level at a higher value than those more concerned with avoiding false rejections. The choice of MCP was not really relevant to these different requirements. For example, they all could have used an STP: the former taking, say, $\alpha = 0.25$ and the latter $\alpha = 0.05$. This would have given everybody the sort of levels and powers he desires, without any need for different MCPs. It is strange that the issue of choice of suitable levels has been so often confused with the technical statistical problem of deciding on a good procedure for multiple comparisons. Could it be that the reason for not taking this simple course is that it does not seem scientifically respectable to work explicitly with a level of 0.25? (See [9, p. 472].)

1.6 Stepwise Procedures for Carrying Out Multiple Comparisons and Some Advantages of STPs

The tests of (1.2) may be carried out in a stepwise manner; the adjective "sequential" [12] is more apt but is already used in a different context in statistics. Instead of looking at the statistics for all sets R containing the set P to be tested, a procedure which involves a great deal of repetition, one begins with the total set and proceeds only to those smaller sets which are not contained in any retained set. (1) The first step is to test the total set K: (a) If $T_K \leq \zeta_k$, ω_K is retained and with it ω_P for all other sets P—hence, no further testing is required and one stops after the first step; (b) if $T_K > \zeta_k$, ω_K is rejected and one proceeds. (2) The second step consists of testing all subsets P of $p = k - 1$ populations: (a) if $T_p \leq \zeta_{k-1}$, P and all its subsets are retained; (b) if $T_P > \zeta_{k-1}$, P is rejected. If all sets P of size $p = k - 2$ are contained in retained sets one stops after the second step. (3) Otherwise, testing continues step after step so long as there remain any untested sets which are not contained in retained sets.

STPs are the only MCPs which ensure coherence without requiring stepwise testing [9, Sec. 9]. This is because both the range and the sum of squares statistics satisfy the monotonicity property

$$T_P \leq T_R \quad \text{if} \quad P \subseteq R \ . \quad (1.21)$$

As a result, for any critical value ζ,

$$T_P > \zeta \quad (1.22)$$

implies $T_R > \zeta$, $\forall R (P \subseteq R \subseteq K)$ so that the rejection criterion (1.2) simply becomes (1.22). In other words, if $T_P > \zeta$, it is impossible for any set R containing P to be retained. Hence, event (1.22) is a sufficient basis for a rejection decision on P and there is no need to refer back to the statistics T_R for larger sets. An STP thus allows all subsets of K to be tested simultaneously without reference to one another and without any stepwise procedure. The only reason one may want to proceed stepwise is that this may obviate the need to compute the statistics for sets contained in larger retained sets. These subsets must necessarily have statistics smaller than ζ and would be retained anyway.

STPs are the only MCPs whose decisions extend beyond hypotheses on subset homogeneity and further allows decisions on all contrasts in the means including simultaneous confidence statements (Tukey and Scheffé bounds). Such decisions and bounds are not possible for the other MCPs, presumably because their decision on a particular set P depends in part on the configuration of the populations not belonging to the set. We find this dependence to be an intuitively undesirable feature of all such MCPs but our feelings may not be shared by others (e.g., Tukey [30]), and especially not by Bayesians.

1.7 Maximal Type I Error Probabilities of MCPs

At this stage, we want to check the maximal probability of any Type I error of each one of the MCPs. It is sufficient to consider false rejection on maximal homogeneous sets, i.e., sets not contained in any homogeneous set. This is because any other homogeneous set P must be a proper subset of some maximal homogeneous set R, and coherence ensures that whenever ω_P is (falsely) rejected so is also ω_R.

As noted in Section 1.3, the probability of false rejection of ω_P reaches its maximum value, the nominal level γ_P, when all true treatment means outside P are widely separated from those of P. Consider, then, a configuration

of the treatments with q widely separated homogeneous sets of true means $P_1, P_2, \cdots, P_q$. For both $T^{(1)}$ and $T^{(2)}$, the statistics $T_{P_1}, T_{P_2}, \cdots, T_{P_q}$ are ratios whose numerators are independent of each other and of the denominator, s or s^2, respectively, common to all of the ratios. Kimball's "improved Bonferroni inequality" [18] therefore ensures that

$$P_{\mathfrak{A}}(\bigcup_{i=1}^{q} (T_{P_i} > \zeta_{p_i})) \leq 1 - \prod_{i=1}^{q} (1 - \gamma_{p_i}) \ . \quad (1.23)$$

In the case of known σ^2, the statistics are actually independent and (1.23) becomes an equality.

Under the Newman-Keuls allocation of levels (1.10) for known σ^2—and, approximately, for large error d.f. n_e—the equality in (1.23) becomes

$$P^{NK}(\text{Type I error}) = 1 - (1 - \alpha)^q \ . \quad (1.24)$$

Clearly, for any $q > 1$

$$P^{NK}(\text{Type I error}) > \alpha \ . \quad (1.25)$$

Under the Ryan allocation (1.12), (1.23) becomes

$$\begin{aligned} P^{R}\ (\text{Type I error}) &\leq 1 - \prod_{i=1}^{q} (1 - \alpha)^{p_i/k} \\ &= 1 - (1 - \alpha)^{\sum_{i=1}^{q} p_i/k} \\ &= 1 - (1 - \alpha) \\ &= \alpha \end{aligned} \quad (1.26)$$

because $\sum_{i=1}^{q} p_i = k$, the number of populations in all the q homogeneous subsets being the total number k. *A fortiori*, under the Duncan allocation and the STP which have lower γ's—(1.18), (1.20)—the Type I error probability never exceeds α.

The possibility that (1.25) may hold renders the Newman-Keuls allocation unsatisfactory unless it is modified to ensure that the false rejection probability be always bounded by the experimentwise level.

1.8 Modifications of Newman-Keuls MCPs

The Newman-Keuls procedure, in its original form, makes a retain-reject decision on every set of means. It may be modified by adding decisions on all collections of disjoint sets, i.e., all partitions, and ensuring coherence through automatic retain decisions on all subsets that belong to a retained partition. Thus, for $k = 5$, the modified procedure would include a nominal α-level test of the simultaneous equality of the first two means and the last three means. If this joint equality hypothesis is retained, so are the separate $\omega_{(1,2)}$ and $\omega_{(3,4,5)}$ hypotheses. If the joint hypothesis is rejected, each of $\omega_{(1,2)}$ and $\omega_{(3,4,5)}$ is tested at nominal α. Clearly, this modification reduces the chance of rejection and prevents the probability of false rejection from exceeding α.

Any choice of nominal α level tests for partitions will determine a modified procedure of this type. Peritz [22] has suggested rejection of the joint hypothesis that all subsets of a partition are homogeneous if any one of these subsets is rejected at the nominal Ryan level (1.12). Thus, for $k = 5$, the joint hypothesis $\omega_{(1,2)} \bigcap \omega_{(3,4,5)}$ is rejected if either $\omega_{(1,2)}$ is significant at $1 - (1 - \alpha)^{2/5}$ or $\omega_{(3,4,5)}$ at $1 - (1 - \alpha)^{3/5}$. If neither is significant, both $\omega_{(1,2)}$ and $\omega_{(3,4,5)}$ are retained without further testing. If either is significant, $\omega_{(1,2)}$ and $\omega_{(3,4,5)}$ are *each* tested at α. (Another procedure for sum of squares statistics can be formulated along the lines of the STP proposed in [11]; its decisions were found to be very similar to those of Peritz's proposal.)

Peritz's modified procedure is a mixture of Ryan's and Newman-Keuls's, leading to fewer rejections than the Newman-Keuls procedure, but somewhat more than Ryan's MCP. Evidently the power must also be larger than that of Ryan's. In fact, it is clear that (1.18), (1.19) and (1.20) remain true for this modified form of the Newman-Keuls technique. Unfortunately, the cross-checking of all subset and partition tests is exceedingly cumbersome. It is included in this study mainly to show how the Newman-Keuls procedure would perform if modified so as not to exceed the experimentwise level.[3]

1.9 Interim Conclusions

In choosing between the different MCPS, one must therefore weigh the greater power of the modified Newman-Keuls, Ryan and Duncan methods against the advantages of STPS: greater computational simplicity (no stepwise procedure is necessary), extension of decisions to all contrasts (and not simply to subsets), availability of corresponding simultaneous confidence bounds and the fact that the decision on any set P is not dependent on the means outside P. The choice will depend on how much larger the power of the stepwise MCPS actually is, which is studied in Section 2.

2. THE MONTE CARLO STUDY

2.1 On Comparing MCPs

A primary problem in evaluating and comparing MCPS is the choice of criteria. A single test is commonly evaluated in terms of its power, and even that becomes difficult when the alternative is composite. An MCP includes a large number of test decisions, each of which has power with respect to every alternative configuration. Which one of them is to be used as a criterion? Moreover, multiple comparisons may be regarded not simply as a collection of retain-reject decisions but also as methods of ordering, or at least partially ordering, the populations under study. To evaluate the operating characteristics of such ordering, one would need more intricate concepts than those of mere power. For the simplest case of pairwise decisions, one already has to add probabilities of errors of Type III, i.e., reversing the true order of a pair. (However, these have been found [2, 15 and 23] generally to be quite small compared to Type II error probabilities.)

[3] A slightly different modification by Welsch [31] has come to our attention recently.

This study is limited to the criterion of power, but considers power for the tests of all subsets of populations. Thus, for example, for $k = 5$ means, we consider powers for the total set, for each of five sets of four treatments, 10 sets of three and 10 sets of two. There is some problem in organizing this large collection of power probabilities, but it turned out to be easier than we had expected (Section 2.4).

To allow meaningful power comparisons, it was necessary to fix the experimentwise level α for all methods used in this study. As noted earlier, this leaves two factors for comparison: (a) the statistics used and (b) the allocation of nominal significance levels to the tests of subsets of different sizes.

This has been restricted to normal populations of equal variance. The statistics used were therefore the range of means and the "between" sum of squares, both in Studentized form. The powers of these statistics had been studied and compared before for overall tests [5]. Our task was to compare them for tests of subsets occurring as part of an MCP.

The main issue for this study was the effect of different allocations of the nominal significance levels. This seemed to be the principal factor confusing the choice between MCPs. On the whole, its effect is likely to be similar for most statistics, so that the limitation to two statistics should not seriously affect this study.

2.2 Issues Requiring Monte Carlo Studies

It is difficult to calculate powers of MCPs because the decision on any set P depends (1.2) not only on the statistic T_P for the set, but also on the statistics for all other sets containing P. The power thus depends on the joint distribution of all these statistics and no one has yet dared work with this. Instead, one tries to estimate power by simulated random experiments—the Monte Carlo method. Note that STPs are an exception to this difficulty since their decisions on P are based exclusively on T_P and powers can be evaluated from the noncentral distributions of the individual statistics involved.

The general discussion of methods of multiple comparisons (Section 1) has allowed us to narrow the issues needing clarification by Monte Carlo study. Our main analytic conclusion (1.20) was that for any one statistic, and fixed experimentwise level, the powers would be (increasing) in the order of the allocations of (i) the STP, (ii) Duncan, (iii) Ryan and (iv) Newman-Keuls, even when modified as suggested by Peritz. The practical problem then is *by how much* these powers actually increase, for it was noted that the STP has various other advantages that one would presumably be willing to forego only for a substantial gain in power. We also noted that the Newman-Keuls allocation is unsatisfactory without modification because of the possibility of very large probabilities of some false rejection, and we therefore have included only its modification.

2.3 Layout of the Monte Carlo Study

The number of normal treatment means studied was $k = 3, 4, 5$ and samples of size 9 were drawn for each treatment. Thirty-four configurations of true treatment means (or expectations) were studied (Table 1). These include configurations $\mu_1 = \cdots = \mu_k = 0$ which provide checks on the true levels of the MCPs. The variance was arbitrarily set at unity.

1. Configurations of True Means $\mu_1, \cdots, \mu_k$ in Monte Carlo Study (Variance Always Unity)

No. of samples k	True means μ_1	μ_2	μ_3	μ_4	μ_5	Noncentrality parameter λ^2	True range
3	0	0	0	—	—	0.0	0
	0	½	1	—	—	4.5	1
	0	0	1	—	—	6.0	1
	0	¾	1½	—	—	10.125	1.5
	0	0	1½	—	—	13.5	1.5
	0	1	2	—	—	18.0	2
	0	0	2	—	—	24.0	2
4	0	0	0	0	—	0.0	0
	0	0	0	½	—	1.6875	0.5
	0	0	½	½	—	2.25	0.5
	0	½	½	1	—	4.5	1
	0	0	½	1	—	6.1875	1
	0	0	0	1	—	6.75	1
	0	0	1	1	—	9.0	1
	0	0	1	1	—	9.0	1
	0	0	1½	1½	—	20.25	1.5
	0	½	1½	2	—	22.5	2
	0	0	2	2	—	36.0	2
5	0	0	0	0	0	0.0	0
	0	0	0	0	½	1.8	0.5
	0	0	0	½	½	2.7	0.5
	0	0	½	½	½	2.7	0.5
	0	0	½	½	1	6.3	1
	0	0	0	½	1	7.2	1
	0	0	0	0	1	7.2	1
	0	1	1	1	1	7.2	1
	0	0	½	1	1	9.0	1
	0	0	1	1	1	10.8	1
	0	0	½	1	1½	15.3	1.5
	0	0	1	1	1½	16.2	1.5
	0	0	0	1	1½	18.0	1.5
	0	0	1	1	2	25.2	2
	0	0	0	1	2	28.8	2
	0	0	0	2	2	43.2	2

NOTE: Some configurations are essentially equivalent to others—but 1,000 separate replications of k samples were run separately for each one.

The configurations are, for each k, given by the order of the true range and true sum of squares—noncentrality parameter

$$\lambda^2 = 9[\sum_{i=1}^{k} \mu_i^2 - (\sum_{i=1}^{k} \mu_i)^2/k] \ . \qquad (2.1)$$

Most of the configurations thus differ in the "variance" λ^2 of the means as well as in the internal arrangement. Thus, e.g., for $k = 4$ and true range 1, Table 1 shows configurations from 0, ½, ½, 1 with $\lambda^2 = 4.5$ to 0, 0, 1, 1 with $\lambda^2 = 9$. These would be the configurations with, respectively, minimal and maximal overall power of the range sum of squares for roughly equal power of the range (see [5]).

2. Monte Carlo Estimates of Power of Four Methods of Multiple Comparisons at $\alpha = 0.05$ for $k = 3$, $n = 9$, with True Means $\underline{\mu}' = (0,0,1)$ and Unit Variance

Method[a]	Subset P							
	(1,2,3)		(1,2)		(1,3)		(2,3)	
	F	Q	F	Q	F	Q	F	Q
	(.521)[b]		(.0148)	(.0189)	(.318)	(.360)	(.318)	(.360)
STP	.535	.525	.016	.022	.347	.384	.332	.370
Duncan	.535	.525	.027	.027	.396	.395	.388	.385
Ryan	.535	.525	.035	.035	.423	.414	.419	.407
Newman-Keuls	.535	.525	.049	.048	.453	.437	.452	.435

[a] F = sum of squares statistics; Q = range statistics.
[b] Values in parentheses are exact powers.

One thousand replications of the k samples of size 9 were simulated for each configuration. (A description of the method of generating these samples can be obtained from the authors.) For each sample, the statistics $T_P^{(1)}$ and $T_P^{(2)}$ were computed and compared with the critical values ζ_p^M for each MCP M as calculated (1.3) from nominal levels γ_p^M based on experimentwise $\alpha = 0.05$. The test decisions for each MCP were sequenced in a stepwise manner—as in Subsection 1.6—so as to simulate the MCPs.

For each set P, the number of replications was counted in which ω_P was rejected. The proportion of such rejections was printed out for each MCP and each set of every one of the 34 configurations. These were our Monte Carlo estimates of the true powers. Examples of all the power estimates for two of the configurations are shown in Tables 2 and 3.

2.4 Summarization of the Monte Carlo Results

To evaluate the large collection of power estimates—for each subset of each configuration, for each method of allocation and both statistics—it was useful to concentrate on comparing powers of each stepwise MCP with that of the corresponding STP. Thus, for each statistic, a plot of each stepwise MCP power against the STP power was made for all sets of a given size $p(= 2, \cdots, k)$ from all configurations of a given size $k(= 3, 4, 5)$. For example, Figure A shows the plot of Ryan's MCP's power against that of the STP, both using range statistics. The points in the plot are for all sets of $p = 3$ treatments from configurations of $k = 5$. Similarly, Figure B shows the corresponding plot for the sum of squares statistics. Altogether nine such plots ($p = 2, \cdots, k$; $k = 3, 4, 5$) were made for each one of the two statistics.

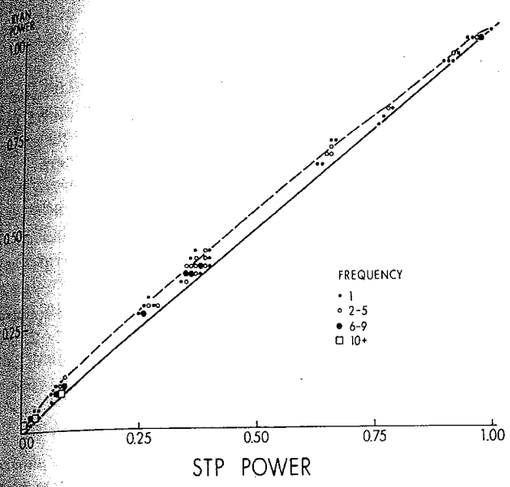
On all plots, the points were found to cluster in a fairly narrow band. Much of this scatter must have been random, considering that each of the coordinates had a standard error of $(\pi(1 - \pi)/1000)^{\frac{1}{2}} \leq (\frac{1}{2} \times \frac{1}{2}/1000)^{\frac{1}{2}} = 0.016$, where π is the true power. However, detailed study of the cases where stepwise power was large relatively to STP power showed these to belong to sets P for which λ_P^2 was appreciably smaller than the overall noncentrality parameter λ_K^2. Evidently, when there were

3. Monte Carlo Estimates of Powers of Methods of Multiple Comparisons at $\alpha = 0.05$ for $k = 4$, $n = 9$, with True Means $\underline{\mu} = (0,0,1.5,1.5)$ and Unit Variance

Method[a]	F	Q	F	Q	F	Q	F	Q	F	Q	F	Q
	Subset P											
	(1,2,3,4)		(1,2,3)		(1,2,4)		(1,3,4)		(2,3,4)		(1,2)	
	(.965)[b]		(.806)		(.806)		(.806)		(.806)		(.006)	(.010)
STP	.956	.917	.794	.814	.789	.799	.812	.822	.783	.800	.003	.009
Duncan	.956	.917	.851	.829	.830	.812	.857	.835	.821	.813	.011	.017
Ryan	.956	.917	.864	.834	.842	.820	.868	.840	.842	.822	.027	.026
Newman-Keuls	.956	.917	.893	.863	.866	.847	.890	.864	.871	.839	.059	.058
Peritz[c]	.956	.917	—	—	—	—	—	—	—	—	.027	.026
	(1,3)		(1,4)		(2,3)		(2,4)		(3,4)			
	(.587)[b]	(.677)	(.587)	(.677)	(.587)	(.677)	(.587)	(.677)	(.006)	(.010)		
STP	.604	.695	.588	.666	.573	.665	.594	.674	.008	.014		
Duncan	.710	.725	.680	.699	.685	.708	.682	.703	.017	.018		
Ryan	.772	.743	.742	.717	.746	.723	.737	.716	.024	.023		
Newman-Keuls	.819	.794	.797	.777	.805	.775	.786	.763	.051	.051		
Peritz[c]	.819	.794	.797	.777	.805	.775	.786	.763	.024	.023		

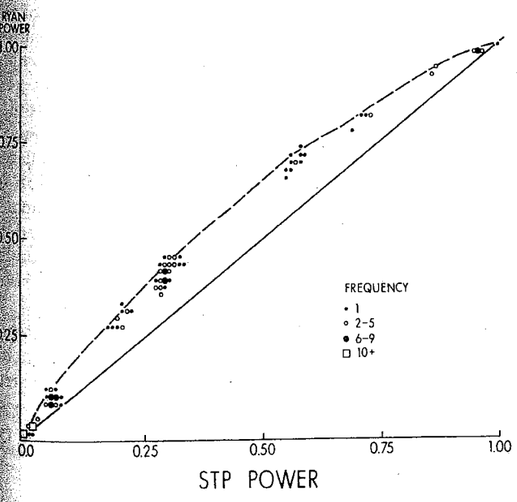
[a] F = sum of squares statistics; Q = range statistics.
[b] Values in parentheses are exact powers.
[c] This is a modification of the Newman-Keuls method.

A. Scatter Diagram of Power of Range Methods: Ryan's Against STP ($k = 5$, $p = 3$)



considerable deviations from the null hypothesis outside P, this enhanced the likelihood that the stepwise MCP ever reached P, and thus increased the power for P. (In the STP, of course, the configuration outside P was irrelevant.) However, these systematic deviations only slightly affected the narrowness of the bands of points on the plots.

B. Scatter Diagram of Power of Sum of Squares Methods: Ryan's Against STP ($k = 5$, $p = 3$)



For a summary description of the relation between stepwise MCP power β^M and STP power β^{STP}, we therefore looked for functions which would fit the band well over its entire range. After some experimentation, we found good fits with

$$\beta^M = (\beta^{STP})^{(1-b)} \tag{2.2}$$

for the range statistics, and with

$$\beta^M = (1 - \beta^{STP})(\beta^{STP})^{(1-a)} + (\beta^{STP})^{(2-c)} \tag{2.3}$$

for the sum of squares statistic. Both these curves pass through

$$\beta^M = \beta^{STP} = 0 \tag{2.4}$$

as well as

$$\beta^M = \beta^{STP} = 1 \ , \tag{2.5}$$

which are necessary conditions for such a relation. The relationship is exponential and the curvature of the exponential over the line $\beta^M = \beta^{STP}$ is measured in (2.2) by a parameter b, $(0 \leq b \leq 1)$. In (2.3) this curvature varies from a to c $(0 \leq a, c < 1)$ as β^{STP} varies from zero to one. Clearly, $\beta^M = \beta^{STP}$ over the entire range if and only if b, a and c are zero. The larger these parameters the higher the exponential above the 45° line $\beta^M = \beta^{STP}$. Least squares[4] estimates of these parameters are given in Tables 4A and 4B along with typical values of β^M calculated for $\beta^{STP} = 0.5$ and 0.75.

Another comparison of interest was that of the powers of the sum of squares Ryan MCP with those of the range STP. These scatterplots were very similar to Figure A, so functions of type (2.2) were fitted as broken lines to each of these plots and the results summarized in Table 5.

2.5 The Monte Carlo Results

The summarized results in Tables 4A and 4B show that the excess of stepwise MCP power over STP power is much greater for the sum of squares than for the range. It also depends on the sizes of the subset and the total set of the configuration—the excess is greatest for the smallest subsets of the largest configurations. And, of course, it depends on the nominal level allocation.

The results for range methods (Table 4A) confirms the order of powers to be

$$\beta^P > \beta^R > \beta^D > \beta^{STP} \ , \tag{2.6}$$

which agrees with (1.20) when Peritz's modification of the Newman-Keuls MCP is used. However, most of the differences are exceedingly small, increasing the power at most (for $k = 5$, $p = 2$) from $\beta^{STP} = 0.5$ to $\beta^P = 0.56$ and from $\beta^{STP} = 0.75$ to $\beta^P = 0.79$. For smaller configurations and/or larger sets the increases are even less.

The results for sums of squares MCPs (Table 4B) show the same order, but the differences are much more noticeable than those for the range. In the extreme case ($k = 5$, $p = 2$) the increase is from $\beta^{STP} = 0.50$ to $\beta^P = 0.72$ and

[4] Fitted by "Marquart's compromise" method of iterative nonlinear least squares [6, Ch. 10] as programmed in IBM Share library program No. 3094.

4. Powers of Various Range and Sum of Squares MCP's in Comparison with STP Power

Method	k	p	Estimate	Power of method	
A. Powers of Various Range MCP's					
			b of (27)	$\beta^{STP} = 0.50$	$\beta^{STP} = 0.75$
Duncan	3	2	.065	.523	.764
Ryan	3	2	.133	.548	.780
Duncan	4	3	.028	.510	.756
Ryan	4	3	.047	.517	.760
Duncan	4	2	.065	.523	.764
Ryan	4	2	.105	.538	.773
Peritz	4	2	.126	.546	.778
Duncan	5	4	.025	.509	.755
Ryan	5	4	.036	.513	.758
Duncan	5	3	.053	.519	.762
Ryan	5	3	.080	.529	.768
Peritz	5	3	.113	.541	.775
Duncan	5	2	.082	.529	.768
Ryan	5	2	.134	.549	.779
Peritz	5	2	.164	.560	.786
B. Powers of Various Sum of Squares MCP's					
			a of (28)	c of (28)	$\beta^{STP} = 0.50$	$\beta^{STP} = 0.75$
Duncan	3	2	.117	.232	.565	.795
Ryan	3	2	.191	.313	.596	.814
Duncan	4	3	.054	.256	.558	.796
Ryan	4	3	.064	.322	.574	.808
Duncan	4	2	.107	.381	.595	.821
Ryan	4	2	.152	.573	.650	.860
Peritz	4	2	.174	.757	.705	.896
Duncan	5	4	.062	.208	.550	.788
Ryan	5	4	.100	.335	.583	.812
Duncan	5	3	.171	.387	.608	.826
Ryan	5	3	.180	.518	.641	.850
Peritz	5	3	.180	.559	.652	.858
Duncan	5	2	.215	.617	.674	.871
Ryan	5	2	.248	.719	.708	.893
Peritz	5	2	.276	.732	.718	.897

from $\beta^{STP} = 0.75$ to $\beta^{P} = 0.90$. Substantial increases also occur for less extreme cases.

With range statistics, power differences between methods are too small to warrant further comment, but not for the sum of squares. Notice first that the modification of the Newman-Keuls MCP mostly has only a very small advantage over Ryan's MCP. Also, the advantage of Ryan's allocation over Duncan's is not large. Essentially the additional power of stepwise MCPs is mostly gained already when the Duncan MCP is employed. The further improvements by the other MCPs contribute relatively small additions to power.

5. Power of Ryan's Sum of Squares MCP in Comparison with Range STP Power

k	p	Estimate b of (27)	Power of Ryan's method	
			$\beta^{STP} = 0.50$	$\beta^{STP} = 0.75$
3	2	.165	.561	.786
4	3	.066	.524	.765
4	2	.110	.539	.774
5	4	.125	.545	.777
5	3	.111	.540	.774
5	2	.126	.546	.778

Comparison of range STPs with sum of squares STPs show the former to have appreciably larger power for small subsets of large configurations, slightly larger power for large subsets and generally slightly smaller power for overall tests.[5] For smaller sets, the extreme observed gain in power is illustrated for $k = 5$, $p = 2$ when $\beta^{STP(Sq)} = 0.50$ corresponds to $\beta^{STP(Range)} = 0.61$. (Similar comparisons of significance levels have been illustrated in [9].)

Similar comparisons for Ryan's allocation shows a slight advantage for the sum of squares when $p > 2$. No difference is evident when $p = 2$. The advantage of the sum of squares powers is at best of the order of 2–3 percent over range powers.

Finally, for comparison of both allocation and statistics, the powers of sums of squares under Ryan's allocation were contrasted with the powers of range STPs (Table 5). The former were found to be slightly larger than the latter, irrespective of size of configuration or subset. The extent of difference can be judged by noting that when range STP powers were $\beta^{STP(Range)} = 0.50$ or 0.75, the corresponding sum of squares Ryan's method had powers of about 0.54 and 0.775, respectively.

2.6 Conclusions and Recommendations

We have argued that Newman-Keuls MCPs are unsatisfactory because their probability of any false rejection may exceed the experimentwise level. We therefore recommend that these MCPs should not be used unless modified to ensure that this probability is bounded by the experimentwise level. Such modifications have been proposed and were found to be the most powerful of the MCPs studied. However, we cannot recommend them for practical use, since they involve impractically complicated procedures and yet provide only very slight gain in power over the next most powerful MCPs. We much prefer Ryan's MCPs, which have almost the same power and seem easy enough to apply in practice, if untabulated percentage points of the F and Studentized range can be obtained. We find no reason to recommend Duncan's MCPs, which entail a slight loss of power compared to Ryan's and are not easier to apply. (The available tables for Duncan's MCPs are for given pairwise comparison levels [14, 15], not for experimentwise levels.)

Against these stepwise MCPs, the STPs, with somewhat smaller power, have the advantages of greater simplicity in operation, of decisions on all contrasts with corre-

[5] The last point agrees with the overall power calculations by David, et al. [5]. These show, for example, that for $k = 5$, $n = 9$ and $\lambda^2 = 13.5$, the sum of squares power 0.663 is towards the upper end of the interval (0.611, 0.668) of the corresponding range powers.

sponding confidence bounds and of excluding extraneous data from the decisions on any set. In case of the range statistics, the gain in power by stepwise MCPS is so small that we feel it cannot possibly outweigh the advantages of the STP. In case of the sum of squares, the power advantage of the stepwise MCPS may be large enough to justify their preference over the STP.

When the choice of statistic is open, the two favored methods are the range STP or the sum of squares Ryan MCP. Direct comparison of the two has shown the latter to have only slightly higher power than the former. Considering all its other advantages, including its associated simultaneous confidence bounds (see [28]), we are led back to recommending the range STP above all other procedures. The only exception would be when there is some commitment to the sum of squares statistic, in which case we would propose its use in Ryan's MCP if only tests are required and in STP form if (see [26]) simultaneous confidence bounds are needed.

27 Limitations

This study has dealt with powers of tests for subset homogeneity among three, four or five normal populations of equal variance when samples of size 9 are taken from each. The questions which remain unanswered are how far our conclusions generalize to other situations. We would guess that changes in sample size and deviations from normality are unlikely to affect our conclusions about the relative merits of different MCPS. We are less certain about the effect of increasing the number of treatments to be compared. Again, it is unlikely that this would affect the order of the powers of the different MCPS, but it might well affect the magnitude of the differences. If so, it might reverse our conclusion about the advantages of the range STP relative to the stepwise MCPS which we found to be only slightly more powerful. This would seem a crucial issue for further Monte Carlo experimentation.

[*Received April 1974. Revised February 1975.*]

REFERENCES

[1] Aitkin, M.A., "Multiple Comparisons in Psychological Experiments," *British Journal of Mathematical and Statistical Psychology*, 22 (November 1969), 193–8.

[2] Balaam, L.N., "Multiple Comparisons—A Sampling Experiment," *Australian Journal of Statistics*, 5 (August 1963), 62–84.

[3] Boardman, T.J. and Moffitt, D.R., "Graphical Monte-Carlo Type I Error Rates for Multiple Comparison Procedures," *Biometrics*, 27 (September 1971), 738–44.

[4] Carmer, S.G. and Swanson, M.R., "Evaluation of Ten Pairwise Multiple Comparison Procedures by Monte-Carlo Methods," *Journal of the American Statistical Association*, 68 (March 1973), 66–74.

[5] David, H.A., Lachenbruch, P.A. and Brandis, H.S., "The Power Function of Range and Studentized Range Tests in Normal Samples," *Biometrika*, 59 (April 1972), 161–8.

[6] Draper, N.R. and Smith, H., *Applied Regression Analysis*, New York: John Wiley and Sons, Inc., 1966.

[7] Duncan, D.B., "A Significance Test for Differences Between Ranked Treatments in an Analysis of Variance," *Virginia Journal of Science*, 2, No. 3 (1951), 171–89.

[8] ———, "Multiple Range and Multiple F-Tests," *Biometrics*, 11 (March 1955), 1–42.

[9] Gabriel, K.R., "A Procedure for Testing the Homogeneity of All Sets of Means in Analysis of Variance," *Biometrics*, 20 (September 1964), 459–77.

[10] ———, "Simultaneous Test Procedures—Some Theory of Multiple Comparisons," *Annals of Mathematical Statistics*, 40 (February 1969), 224–50.

[11] ——— and Sokal, R.R., "A New Statistical Approach to Geographic Variation Analysis," *Systematic Zoology*, 18 (September 1969), 259–78.

[12] Games, P.A., "Multiple Comparisons of Means," *American Educational Research Journal*, 8 (May 1971), 531–65.

[13] Harter, H.L., "Error Rates and Sample Sizes for Range Tests in Multiple Comparisons," *Biometrics*, 13 (December 1957), 511–36.

[14] ———, *Order Statistics and Their Use in Testing and Estimation, Vol. I*, Washington: U.S. Government Printing Office, 1970.

[15] ———, Clemm, D.S. and Guthrie, E.H., *The Probability Integrals of the Range and of the Studentized Range, Vol. II*, Technical Report 58–484, Wright Air Development Center, 1959.

[16] Hartley, H.O., "Some Recent Developments in Analysis of Variance," *Communication in Pure and Applied Mathematics*, 8 (February 1955), 47–72.

[17] Keuls, M., "The Use of the 'Studentized Range' in Connection with an Analysis of Variance," *Euphytica*, 1 (1952), 112–22.

[18] Kimball, A. W., "On Dependent Tests of Significance in the Analysis of Variance," *Annals of Mathematical Statistics*, 22 (December 1951), 600–2.

[19] Kleijnen, J.P.C., *Statistics and Simulation*, to appear.

[20] Miller, R.G., Jr., *Simultaneous Statistical Inference*, New York: McGraw-Hill Book Co., 1966.

[21] Newman, D., "The Distribution of the Range in Samples from the Normal Population, Expressed in Terms of an Independent Estimate of Standard Deviation," *Biometrika*, 31 (July 1939), 20–30.

[22] Peritz, E., "A Note on Multiple Comparisons," unpublished paper, Hebrew University, 1970.

[23] Petrinovich, L.F. and Hardyck, C.D., "Error Rates for Multiple Comparison Methods," *Psychological Bulletin*, 71 (January 1969), 43–54.

[24] Ryan, T.A., "Multiple Comparisons in Psychological Research," *Psychological Bulletin*, 56 (January 1959), 26–47.

[25] ———, "Significance Tests for Multiple Comparison of Proportions, Variances and Other Statistics," *Psychological Bulletin*, 57 (July 1960), 318–28.

[26] Scheffé, H., "A Method for Judging All Contrasts in the Analysis of Variance," *Biometrika*, 40 (June 1953), 87–104.

[27] Spjøtvoll, E. and Stoline, M.R., "An Extension of the *t*-Method of Multiple Comparison to Include the Cases with Unequal Sample Sizes," *Journal of the American Statistical Association*, 68 (December 1973), 975–8.

[28] Tukey, J.W., "Quick and Dirty Methods in Statistics, Part II: Simple Analyses for Standard Designs," *Proceedings of the Fifth Annual Convention of the American Society for Quality Control* (1951), 189–97.

[29] ———, *The Problem of Multiple Comparisons*, unpublished paper, Princeton University, 1953.

[30] ———, Personal communication, 1967.

[31] Welsch, R.E., "A Modification of the Newman-Keuls Procedure for Multiple Comparisons," M.I.T. School of Management Working Paper, 1972, 612–72.