DOCUMENT RESUME

ED 096 356 95 TM 003 972

AUTHOR Hopkins, Kenneth D.; Anderson, Beverly

TITLE Instructional Module on Multiple Comparison Techniques in Research: A Guide for Selecting the "Method of Choice".

INSTITUTION Colorado Univ., Boulder. Lab. of Educational Research.

SPONS AGENCY National Center for Educational Research and Development (DHEW/OE), Washington, D.C.

PUB DATE Sep 73

NOTE 40p.; For related documents, see TM 003 967-973

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE

DESCRIPTORS Achievement Tests; Autoinstructional Aids; *Comparative Analysis; Educational Researchers; Graduate Students; Guides; *Hypothesis Testing; Research Design; *Research Methodology; *Statistics

ABSTRACT

The purpose of this module is to provide the researcher and the consumers of such research with a recognition of the assumptions and appropriate use and interpretation of each of 10 multiple comparison (MC) statistical techniques. In this self-contained and self-instructional module, the user is sensitized to the serious consequences of inappropriate multiple comparison use by employing all MC methods to the same data. He then is introduced to the criteria for selecting the best MC method for a given purpose. Computational considerations follow with self-instructional exercises and mastery tests. A familiarity with the t-test and one-factor analysis of variance is required. (Author/SE)

INSTRUCTIONAL MODULE ON MULTIPLE COMPARISON TECHNIQUES IN RESEARCH

I. A Guide for Selecting the "Method of Choice"

Kenneth D. Hopkins
Beverly Anderson
Laboratory of Educational Research
University of Colorado

September, 1973

NCERD Reporting Form — Developmental Products

1. Name of Product
Instructional Module on Multiple Comparison Techniques in Research.

2. Laboratory or Center
(LER)

3. Report Preparation
Date prepared 11/9/73
Reviewed by K.D. Hopkins, director

4. Problem: *Description of the educational problem this product designed to solve.*

Many research studies are inappropriate and inefficient statistical procedures for comparing three or more groups. Applied textbooks are not adequate in providing the needed competencies for selection of the most powerful multiple comparison (MC) method that will answer the researcher's questions.

5. Strategy: *The general strategy selected for the solution of the problem above.*

The training strategy is: (1) to survey current experimental statistics textbooks to illustrate the characteristicly uneven and inappropriate coverage; (2) contrast the grossly different conclusions that will result depending on MC method used; (3) to provide a flowchart guide to selecte the MC method of choice; (4) to provide self-instructional exercises for developing needed user competencies.

6. Release Date: *Approximate date product was (or will be) ready for release to next agency.*

12/1/73

7. Level of Development: *Characteristic level (or projected level) of development of product at time of release. Check one.*

- ☐ *Ready for critical review and for preparation for Field Test (i.e. prototype materials)*
- ☒ *Ready for Field Test*
- ☐ *Ready for publisher modification*
- ☐ *Ready for general dissemination/ diffusion*

8. Next Agency: *Agency to whom product was (or will be) released for further development diffusion.*

NIE

9. Product Description: *Describe the following; number each description.*

- ● 1. *Characteristics of the product.*
- ● 2. *How it works.*
- ● 3. *What it is intended to do.*
- ● 4. *Associated products, if any.*
- ● 5. *Special conditions, time, training, equipment and/or other requirements for its use.*

Characteristics of the Product:

A 34 page discussion of various MC procedures -- their assumptions, consequences, and interpretations. The module is self-contained and self-instructional. (See also "5. Strategy")

How it Works:

The user is sensitized to the serious consequences of inappropriate MC use by employing all MC methods to the same data. He then is introduced to the criteria for selecting the best MC method for a given purpose. Computational considerations follow with self-instructional exercises and mastery tests.

What it is Intended to do:

Provide the research producer and consumer with a recognition of the assumptions and appropriate use and interpretation of each ten MC techniques.

Requirements for Use:

A familiarity with the t-test and one-factor analysis of variance.

10. Product Users: *Those individuals or groups expected to use the product.*

The product is intended to be used by applied researchers in education and by students in intermediate courses in statistics or experimental design.

11. Product Outcomes: *The changes in user behavior, attitudes, efficiency, etc. resulting from product use, as supported by data. Please cite relevant support documents. If claims for the product are not yet supported by empirical evidence please so indicate.*

Of twenty-eight users responding to anonymous evaluation forms, 46% rated the materials as "very good"; 39% rated them as "good"; and only 14% rated them as "fair" or "poor." The median error rated was 7.5%.

To the question, "Are the materials superflous, i.e., are there other sources that accomplish the same purposes that are as good or better?", 85% of the responses were "No."

The rating of "good" or "very good" by 85% of the users suggests instructional value.

12. Potential Educational Consequences: *Discuss not only the theoretical (i.e. conceivable) implications of your product but also the more probable implications of your product, especially over the next decade.*

The use of this product is expected to result in more appropriate use and interpretation of research studies involving three or more groups.

13. Product Elements: *List the elements which constitute the product.*	**14. Origin:** *Circle the most appropriate letter.*
One self-contained and self-instructional module	[✓ D] M A
	D M A
	D M A
	D M A
	D M A
	D M A
	D M A
	D M A
	D M A
	D M A
	D M A
	D M A
	D M A
	D M A
	D M A

	D= Developed
	M= Modified
	A= Adopted

15. Start-up Costs: *Total expected costs to procure, install and initiate use of the product.*

Reproduction cost only.

16. Operating Costs: *Projected costs for continuing use of product after initial adoption and installation (i.e., fees, consumable supplies, special staff, training, etc.).*

Reproduction.

17. Likely Market: *What is the likely market for this product? Consider the size and type of the user group; number of possible substitute (competitor) products on the market; and the likely availability of funds to purchase product by (for) the product user group.*

Research and evaluation personnel, especially those being trained on the job.

Students in intermediate statistics and experimental design courses.

INSTRUCTIONAL MODULE ON MULTIPLE COMPARISON TECHNIQUES IN RESEARCH

I. A Guide for Selecting The "Method of Choice"[a]

This module has two major components, the first deals with the particular advantages and disadvantages of each, the second presents computational interrelationships of the various procedures.

The need for a researcher's guide to the use of multiple comparison (MC) techniques is illustrated by recent studies by Tringo (1970) and Wilson (1971). Although these are not poor studies, they illustrate the two extremes in their selection and use of a MC technique. Tringo (1970) used multiple t-tests to make comparisons among seven groups; the multiple t-tests produced has an inordinately high risk of falsely rejecting a true null hypothesis. Wilson (1971) employed the Scheffe test to detect significant differences among three means; this method is the most conservative and least powerful of all MC methods for contrasting pairs of means.

When there are more than two treatment or comparison groups being studied, the analysis of variance (ANOVA) or covariance (ANCOVA) will determine whether

[a] Adapted from a forthcoming article in the Journal of Special Education.

or not the differences among means are greater than expected from chance alone. ANOVA or ANCOVA does not, however, proceed to the next logical step of identifying which differences among the means are significant; this is the task of multiple comparison techniques.

Multiple comparison techniques are a relatively recent development in the area of statistical analysis which have direct applicability in behavioral research. Dissemination via applied statistics textbooks has reflected the expected theory-to-practice lag and, in the main, the information exchange has been based more in inertia and precedent than actual research utility.

The lack of systematic textbook coverage of MC methods is illustrated in Figure 1 which given the methods covered by popular applied statistics or experimental design textbooks. Notice that the Scheffe method is the MC technique most commonly treated, yet it is the least powerful MC procedure for responding to typical research questions.

Multiple comparisons are a not-closely-related family of techniques except that they serve a common purpose. This diversity no doubt has contributed to the uneven textbook coverage. Whereas there is a major pathway that leads the learner through the analysis of variance, when he encounters the domain of multiple comparisons, the pathway branches into a network of numerous unmarked routes. Each MC method has unique advantages and disadvantages. Ideally, the researcher should be familiar with the major alternatives so that the method can be selected that yields maximum power for the questions, i.e., so that "the method of choice" will be chosen. In addition, this information is useful in interpreting published research.

All too frequently, the MC technique employed in a study is one with which the researcher s familiar because it "happened" to be treated in the researcher's favorite reference. As a consequence, inappropriate, weak, or at least

Table 1

Coverage[a] of Multiple-Comparison Techniques in Selected Textbooks on Statistics and Experimental Design

Textbook Author(s)[b]	Technique[c]								
	Duncan	Dunn	Dunnett	Marascuilo	Multiple t(LSD)	Newman-Keuls	Planned Orthogonal Comparisons	Scheffe	Tukey
Bailey (1971)							X	X	
Brownlee (1965)								X	X
Dayton (1970)	X		X		X	X	X		
DuBois (1965)									
Edwards (1967)							X	X	
Edwards (1968)	X		X				X	X	
Edwards (1969)								X	
Ferguson (1971)	X					X	X	X	X
Fryer (1966)	X				X				
Glass & Stanley (1970)								X	X
Guilford (1965)									X
Hays (1963)							X	X	
Hays & Winkler (1971)									
Kirk (1968)	X	X	X		X	X	X	X	X
Li (1964)			X				X	X	X
Marascuilo (1971)				X			X	X	X
McNemar (1969)								X	
Myers (1972)			X			X	X	X	X
Ostle (1963)					X			X	
Roscoe (1969)			X					X	
Snedecor & Cochran (1967)					X	X		X	X
Steel & Torrie (1960)	X		X		X	X	X		X
Walker & Lev (1969)								X	
Winer (1971)	X		X		X	X	X	X	X

[a] Mere mention of a technique in a textbook was not considered coverage.

[b] Complete specification of each textbook is included in the references at the end of this paper.

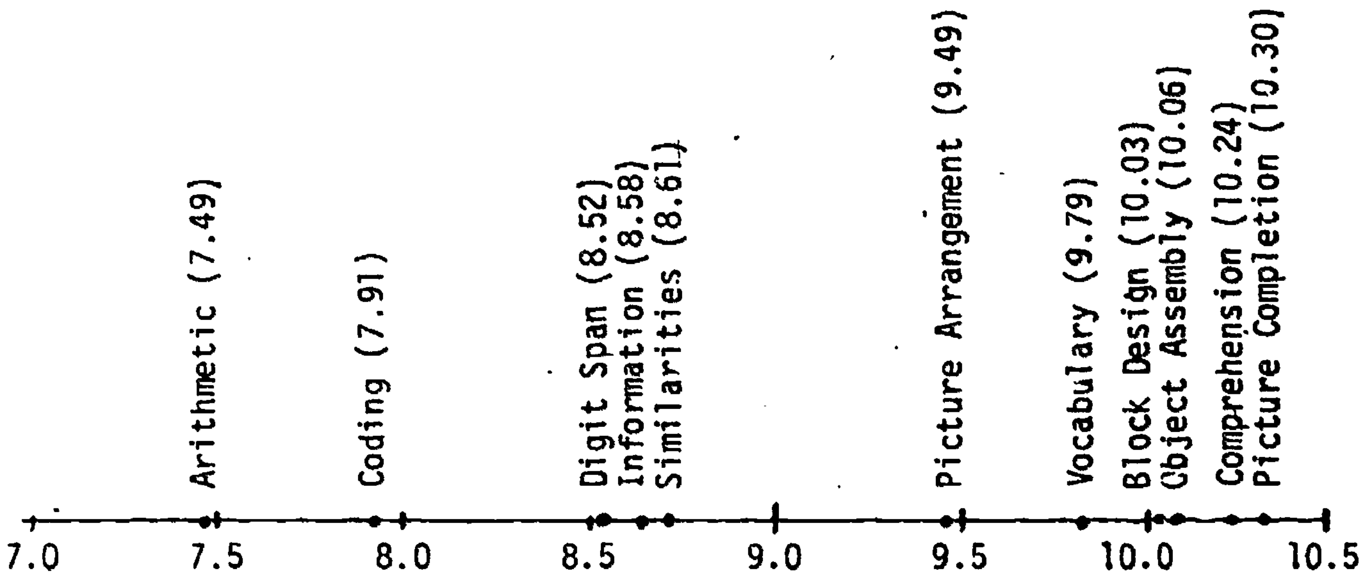
[c] An X indicates coverage of a technique in a specified textbook.

inefficient methods of analysis are frequently used. *Differences in the conclusions reached in a given study can vary markedly depending on the MC technique employed*. In the derivations of the MC methods, different assumptions and restrictions are imposed. As a general rule, the more limitations the researcher can live with, the more powerful will be the statistical tests for the hypotheses of interest if the proper MC alternative is chosen.

The differences among the various multiple comparison techniques will be illustrated from an actual study (Hopkins, 1964) that examined the pattern of performance of 33 diagnosed neurologically handicapped children (ages 6-12) on eleven subtests of the *Wechsler Intelligence Scale for Children* (WISC). The results of the Subtests-by-Subjects analysis of variance revealed a highly significant difference among subtest means ($F = 32.92/6.99 = 4.71$, $p < .001$). The subtest means are graphically presented in Figure 2.

To illustrate the great variation in conclusions as a consequence of the MC techniques employed, all possible differences in pairs of means were tested for significance using the various MC alternatives: multiple t-test, Duncan's New Multiple Range Test, Newman-Keuls test, Tukey test, Dunn test, Marascuilo test, and Scheffe test. Of the possible 55 comparisons of pairs of means, the number of significant mean differences at the .05 and .01 levels for each method differs greatly as is shown in Table 1. For example, with $\alpha = .05$, the number of null hypotheses rejected varied from 1 using the Scheffe to 24 for the Duncan and multiple t-tests; with $\alpha = .01$, the number of significant differences in means varied from 0 to 15. How can such inconsistency in conclusions result from the use of alternative MC approaches?

α-Considerations. Even though each method has the same nominal α-value, not all are appropriate in this situation. Although commonly used, the multiple-t approach is never the method of choice and cannot be recommended. Multiple



Fiqure 1

WISC Subtest Means of 33 Neurolrgically Handicapped Children.

t-tests (also known as the "least significant difference" or lsd procedure) introduces an inextricable pattern of dependency, and yields inaccurate probability statements regarding the null hypotheses. The inaccuracy is magnified in direct proportion to the number of means in the set being examined. In the present example with 11 means, 55 different t-tests would be required to test all combinations of pairs. Even if all pair-wise null hypotheses were true, more likely than not, the lowest vs. the highest mean from the 11 subtests would yield a t-ratio that would be ruled "significant" at the .05 level.

The Duncan method has the peculiar property of using a fluctuating α-rate depending on the number of means in the set being examined. The true probability of a type-I error (rejecting a true null hypothesis) is always larger than the tabled α-value except when there are only two means in the set being tested. For this reason the authors view the Duncan procedure as never the method of choice, in spite of its popularity. For example, if the Duncan method was used to test $H_0: \mu_1 = \mu_{11}$ (the smallest and largest means in our sample), the critical value for the α = .05 value in Duncan's table, will be exceeded 40% of the time even when the null hypothesis is true (almost as often as with the multiple t approach). In other words the true probability of a type-I error (incorrectly rejecting a true null hypothesis) is not what most users naturally assume, e.g., .05, but much larger -- .40 in our example.

The remaining techniques given in Table 1 have accurate α-values, but α in relation to what? In the Newman-Keuls method, α is .05 for each individual null hypothesis (H_0) tested, i.e., a contrast based error rate. In the Dunn, Dunnett, Tukey, Marascuilo, and Scheffe methods, α is .05 for the entire set or family of H_0's to be tested in the experiment; i.e., an experiment based error rate.

Table 1

Number of Significant Differences (of the 55 possible) Between Pairs of WISC Subtest Means for Various Multiple Comparison Methods

MC Method	Number of H_0's Rejected		Percent of H_0's Rejected	
	@ α = .05	@ α = .01	@ α = .05	@ α = .01
Multiple t (LSD)[a]	24	15	44%	27%
Duncan[a]	24	11	44%	20%
Newman-Keuls	11	6	20%	11%
Tukey	9	4	16%	7%
Dunn	7	4	13%	7%
Marascuilo	3	1	5%	2%
Scheffé	1	0	2%	0%

[a] For these methods the actual probability of a type-I error is considerably greater than the tabled, nominal α-value.

The Newman-Keuls method will tend to reject more pair-wise H_0's than the other accurate (with respect to type-I error probabilities) methods because of its differently based error rate. It should be noted however, that the critical value for Tukey and Newman-Keuls methods will always be equal when testing the extreme-most means, i.e., when H_0: $\mu_{smallest} = \mu_{largest}$ is being tested; hence they will always lead to the same conclusion for this H_0. Thus, although the Newman-Keuls procedure has a contrast based error rate, in this limited sense the Newman-Keuls method has an experiment based error rate, i.e., it, and the Tukey method, will be expected to make a type-I error when testing the extreme-most means in 5% of the experiments in which all pair-wise H_0's are true. However, in these %5 of the experiments, when going on to test other pairs of means the Newman-Keuls method will tend to make more type-I errors than will the Tukey procedure.

Of the common procedures, the Scheffe method is the least powerful for detecting differences between pairs of means. It is best, however, for data snooping and testing complex hypotheses (hypotheses involving more than two means). The Marascuilo (1966) method is a rather recently devised MC procedure appropriate for studies employing large samples. Unlike the others it does not assume homogeneity of variance, but does require large samples. It is more useful for making multiple comparisons among correlation coefficients and proportions.

The relative power and sensitivity of the various MC alternatives are also illustrated in Figure 3. In this figure, the magnitudes of minimum differences between the WISC subtest means required for significance for the various MC methods are graphically depicted. Equivalently, the relative magnitudes for the associated confidence intervals are illustrated in Figure 3 (except for the Duncan and Newman-Keuls methods which do not lend themselves to interval

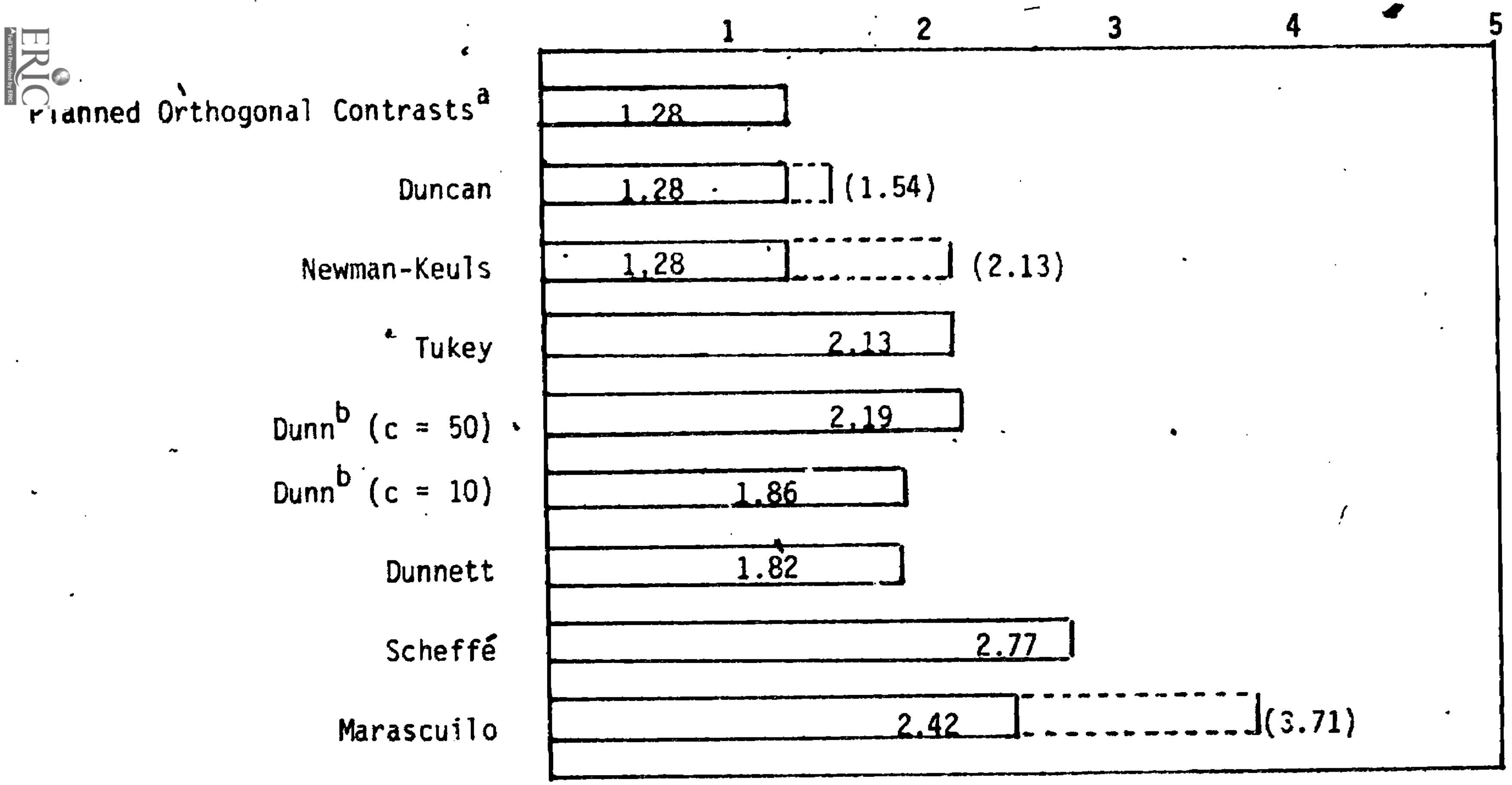


Figure 2

Difference in WISC Subtest Means Required for Significance in Selected Multiple-Comparison Techniques[c]

[a] Planned orthogonal contrasts have the same value as multiple *t* tests(LSD), but without the inflated α-values.

[b] The Dunn procedure requires that the number of contrasts be set a priori. In this example there are 55 possible comparisons between pairs of means, hence the nearest tabled value, c = 50, was selected. The value for 10 planned comparisons was also included to allow comparison of the sensitivity of the Dunn with the Dunnett method which allows 10 contrasts in the illustrated situation.

[c] The values given are the magnitudes of $\underline{M}_i - \underline{M}_j$ needed to reject pairwise $\underline{H}_o$'s for various MC methods using Hopkins (1964) data ($\underline{MS}_{error}$ = 6.99, $\underline{J}$ = 11, $\underline{df}_{\underline{e}}$ =320, and $\underline{\alpha}$ = .05).

For those methods not requiring a constant difference for all pairs of means (Duncan, Newman-Keuls, and Marascuilo), the greatest (represented by dotted lines and values in parentheses) and least differences are indicated.

estimation). For those methods which do not have a single critical value required for all mean differences, the greatest and least values are given. For example, for the largest difference between pairs of means, (i.e., the difference between the Arithmetic and Picture Completion means as illustrated in Figure 2) a value of 2.13 is required to reject the null hypothesis for both the Tukey and Newman-Keuls methods, yet the latter requires a mean difference of only 1.28 for adjacently ordered means.

Clearly such disparity in results is undesirable, but how does one go about selecting the optimum procedure for a given research study? Figure 4, which is a revision of an early schema (Hopkins and Chadbourn, 1967), gives a flow chart to illustrate the critical decisions leading to the method of choice in a given research situation.

Criteria for Selecting a Multiple Comparison Method

Since the treatment of multiple comparisons is scattered among many sources, the flow chart given in Figure 4 is provided to assist the researcher in the selection of an appropriate method for use in examining differences between means when more than two groups are involved. In words, the schema illustrates the following decisions.

1. All methods except Marascuilo's[1] assume homogeneity of variances; this assuption should be tested, since unlike ANOVA these procedures do not appear to be robust to non-homogeneity of variances (Petrinovich and Hardyck, 1969), especially with unequal sample sizes.

[1] The large sample method described by Marascuilo (1966) is needed when making multiple comparisons among correlation coefficients, proportions and contingency tables, and is recommended for contrasting means only when variances are not homogeneous.

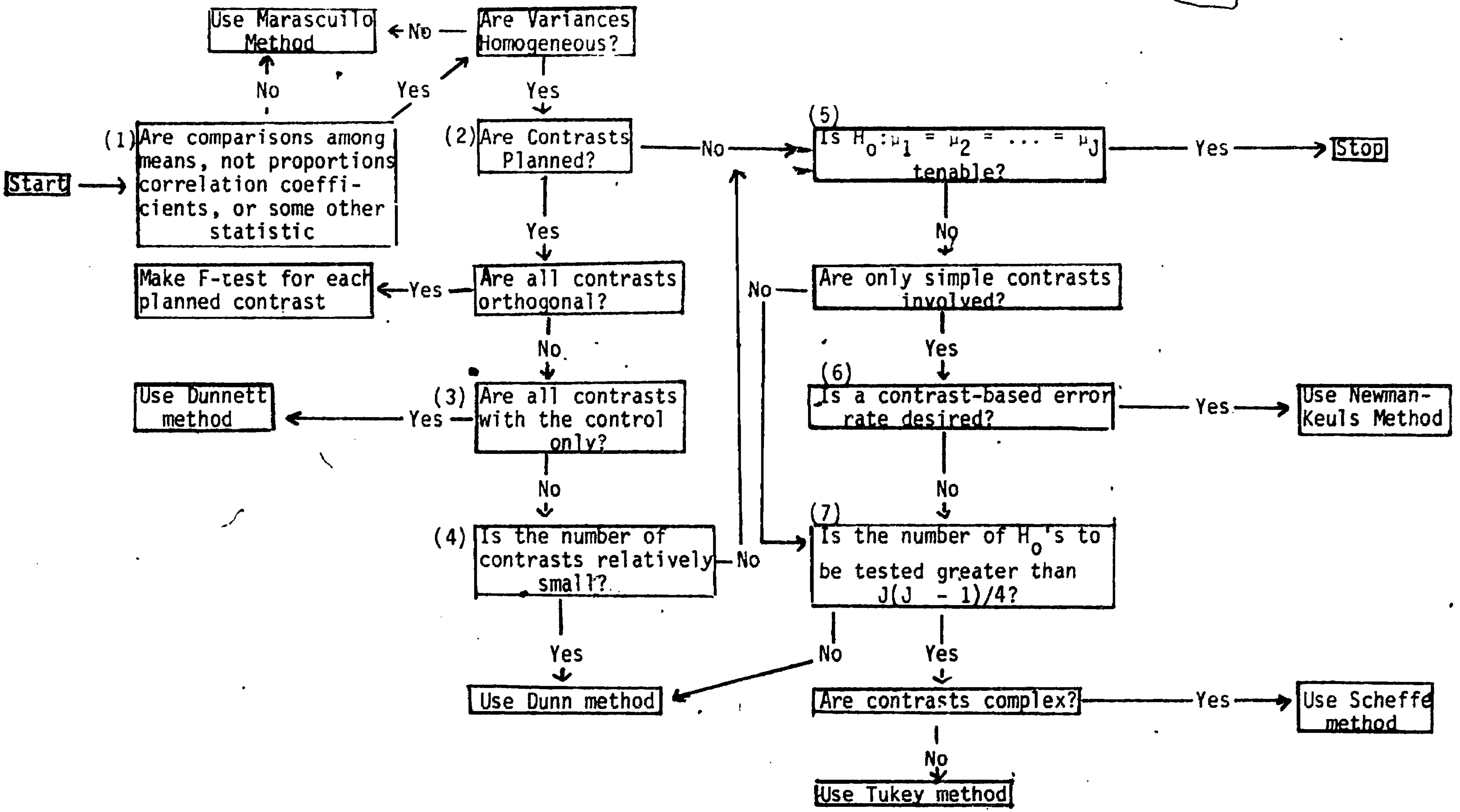

Start
(1) Are comparisons among means, not proportions correlation coefficients, or some other statistic
No
Use Marascuilo Method
Yes
Are Variances Homogeneous?
No
Yes
(2) Are Contrasts Planned?
No
Yes
Are all contrasts orthogonal?
Yes
Make F-test for each planned contrast
No
(3) Are all contrasts with the control only?
Yes
Use Dunnett method
No
(4) Is the number of contrasts relatively small?
No
Yes
Use Dunn method
(5) Is $H_0: \mu_1 = \mu_2 = \dots = \mu_J$ tenable?
Yes
Stop
No
Are only simple contrasts involved?
No
Yes
(6) Is a contrast-based error rate desired?
Yes
Use Newman-Keuls Method
No
(7) Is the number of H_0's to be tested greater than J(J - 1)/4?
No
Yes
Are contrasts complex?
Yes
Use Scheffe method
No
Use Tukey method

Figure 3

A Schema to Guide for the Selection of Multiple Comparison Techniques.

2. Make planned orthogonal contrasts if they will answer the relevant hypotheses (usually this will not be the case). Each comparison would have the contrast as the base for α (see 3 below). The setting of α should not be arbitrary, but influenced by power considerations (Hopkins, 1972).
3. If all comparisons of interest pit the control group against each of the other J-1 groups, use the Dunnett procedure. Using the Dunnett technique, the probability of a type-I error is α for the set of J - 1 tests, i.e., an experiment-based α-value.
4. If the number of comparisons is relatively few, (e.g., 2(J - 2) or less), use the Dunn test. The Dunn test is appropriate for simple (involving only two means, i.e., a pair of means) and complex (involving more than two means), and has an experiment-based α-value.
5. Compare F-ratio differences among the means (obtained in the ANOVA or ANCOVA) with critical value required for significance. If H_0 cannot be rejected, one probably should not look further for mean differences, although this is a logical rather than a purely statistical consideration. If the omnibus F is not significant, it is tantamont to concluding all differences among all means is attributable to random sampling error.
6. Select the base of α (contrast or experiment). The Tukey, Scheffe, Dunn, Dunnett, and Marascuilo MC tests use the experiment as base, hence a type-I error will be made in only 5% of the experiments (if α = .05). The Newman-Keuls method employs the comparison as the unit, therefore, a type-I error can be expected for 5% of the contrasts. This is equivalent to saying more type-I errors for differences between pairs of means will be made with the Newman-Keuls procedure, but fewer type-II errors than with the experiment-based methods. Hence, if only pair-wise comparisons are

involved and the contrast is the base for α, use the Newman-Keuls method.[2] (For unequal sample sizes for the Newman-Keuls or Tukey methods see Steel and Torrie, 1960, p. 114 or Fryer, 1966, p. 274).

7. If the number of hypotheses to be tested is less than J(J - 1)/4, the Dunn (1961) method will usually be more powerful than either the Tukey or the Scheffe method. (Tables 4-6 in Dunn's (1961) article provide precise figures for various J, α, and df_e combinations for which the Dunn methods would be more powerful). The special tables of critical values for the Dunn test are available in Dunn (1961), Miller (1966), Kirk (1968), and Myers (1972). If all J(J - 1)/2 pairwise comparisons are of interest, as is usually the case, the Tukey method should be used since it is more powerful than the Dunn and Scheffe methods under such conditions (Scheffe, 1959, p. 76).

8. If comparisons between complex combinations of means are desired, the Scheffe method has more power than the Tukey.

The most rigorous and comprehensive treatment of the statistical properties underlying multiple comparison procedures is found in Miller (1966). The reader will find quite complete treatments in Kirk (1968), and Winer (1971). Articles by Duncan (1965) and Sparks (1963) provide useful computational comparisons. If one is doing multiple comparisons following an ANCOVA it is important to remember that adjustments must be made in the mean square error term (e.g., see Winer, 1971, p. 772).

The purpose of this article was to illustrate the importance of selecting the appropriate statistical model that best fits the experimental methods and hypotheses of interest. The schema provided was designed to encourage the reader

[2] Duncan's New Multiple Range Test is not included here since the experimenter-selected α-value is correct only for adjacently-ordered means; the actual α-value always exceeds the selected value in all other contrasts (cf. Edwards, 1968, p. 134-135). In addition, mathematical statisticians are not in agreement regarding the validity of certain assumptions employed in its derivation (Scheffe, 1959, p. 78; Duncan, 1965, p. 178).

to consider critical factors that will determine the selection of the optimum multiple comparison method for the hypotheses of interest in a given study.

Instructional Exercises

Which multiple comparison technique is preferable:

1. for testing several correlation coefficients for significant differences?

   ____

   Marascuilo

2. for comparing each of several means with the mean of the control group?

   ____

   Dunnett

3. when, although there are ten treatment groups, only twelve hypotheses are to be tested?________________

   Dunn

4. for making all possible pairwise contrasts among means with a contrast-based error rate? ________________

   Newman-Keuls

5. for making all possible pairwise contrasts among means with an experiment-based error rate? _______________

   Tukey

6. for data snooping -- making post hoc complex contrasts involving means?

   ____

   Scheffe

7. for comparing means when variances are extremely heterogeneous? ____________

   Marascuilo

II. Multiple Comparisons -- Computation

Multiple comparisons are a loosely-related family of techniques for identifying significant differences among a set of three or more means. There are eight principal methods but only three different computational procedures; three employ the t-tests (multiple t, Dunn (or Bonferroni), and Dunnett); three use the studentized range statistic, q, (Tukey, Newman-Keuls, and Duncan); and two employ the F-statistic (Scheffe and planned orthogonal contrasts). In the discussion to follow, it is assumed that the usual ANOVA assumptions hold and that all means are based on the same number (n) of observations.

The t-statistics Approaches

The multiple t, Dunn, and Dunnett methods are computationally identical (for a given H_0), except that the critical t-values required to reject H_0 will differ. The amount of difference is highly related to the number of means, J, being compared. (If J = 2, all methods give identical results, but, of course, are unnecessary.)

Suppose there are six groups of 11 subjects each that are compared on some measure. The analysis of variance (ANOVA) revealed that $H_0: \mu_1 = \mu_2 = \ldots = \mu_6$ is not tenable, and hence rejected. The ANOVA table is given below:

Source of Variation	df	MS	F
Treatments	5	176	8.0
Error	60	22	

But which $H_0: \mu_i = \mu_j$ are tenable? To test each H_0, compute the t-ratio.

$$t = \frac{\bar{X}_i - \bar{X}_j}{s_{\bar{X}-\bar{X}}} = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{2MS_e}{n}}}$$

For simplicity, select $\bar{X}_i$ to be larger than $\bar{X}_j$. MS_e is the error term from the analysis of variance and n is the number of observations on which each mean is based.

In this example: $t = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{2(22)}{11}}} = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{4}} = \frac{\bar{X}_i - \bar{X}_j}{2}$

Is the t-value large enough to be significant, i.e., to reject H_0? The non-directional H_0 is rejected at the α-level if the observed t-value exceeds the critical t-values shown below.

	For Multiple-t	For Dunn[a]	For Dunnett
General expression:	$_{1-\alpha/2}t_{f_e}$	$_{1-\alpha}t_{c,f_e}$	$_{1-\alpha/2}t_{J,f_e}$
or, in our example with α = .05:	$_{.975}t_{60} = 2.00$	$_{.95}t_{c,60} = ?$ $_{.95}t_{5,60} = 2.66$ $_{.95}t_{15,60} = 3.06$	$_{.975}t_{6,60} = 2.63$
Number of H_0's to be tested:	$\frac{J(J - 1)}{2}$	c	J - 1

[a] Note Dunn table presupposed that small value is subtracted from larger, hence the tabled .95 values are actually the .975 point in the cumulative distribution.)

Although the three approaches arrive at identical t-values for a given contrast, they will usually differ greatly in the critical t-values needed to reject H_0. The multiple-t, although widely used will result in many type-I errors (i.e., rejecting true H_0's) and is never recommended as the method of choice.

The Dunnett is appropriate only when one wishes to compare each of the J - 1 groups with one other predesigned groups -- usually the control group. In most instances the researcher wishes to compare each mean with every other mean, hence the Dunnett rarely addresses many of the investigator's questions.

The Dunn test requires that the researcher have planned in advance which comparisons he is going to make. The number of these planned contrasts, c, affects the critical t-value as would be expected -- the larger the value of c, the larger the critical value of t. In our example, the critical t-values are 2.66 and 3.06 for c = 5 and c = 15 respectively. If the researcher wishes to test all J(J - 1)/2 pairwise comparisons (15 in our example), the Dunn procedure is not as powerful as other alternatives to be considered later.

Studentized Range (q) Methods

The Tukey, Newman-Keuls, and Duncan methods are computationally identical except that the critical q-values for a given H_0 will usually differ. The studentized range statistic, q, is:

$$q = \frac{\bar{X}_i - \bar{X}_j}{s_{\bar{X}}} = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{MS_e}{n}}}$$

In the example:

$$q = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{22}{11}}} = \frac{\bar{X}_i - \bar{X}_j}{.2}$$

Is the q-value large enough to be significant, i.e., to reject H_0? The non-directional $H_0: \mu_i = \mu_j$ is related at the α-levels if the observed q-value exceeds the critical q-values given below.

	For Tukey	For Newman-Keuls	For Duncan
General expression:	$_{1-\alpha}q_{J,f_e}$	$_{1-\alpha}q_{r,f_e}$ where r is the number of means in the subset being evaluated	$_{1-"\alpha"}q_{r,f_e}$
or, in our example with α = .05:	$_{.95}q_{6,60} = 4.16$	$_{.95}q_{6,60} = 4.16$	$_{".95"}q_{6,60} = 3.19$
		$_{.95}q_{5,60} = 3.98$	$_{".95"}q_{5,60} = 3.14$
		$_{.95}q_{4,60} = 3.74$	$_{".95"}q_{4,60} = 3.07$
		$_{.95}q_{3,60} = 3.40$	$_{".95"}q_{3,60} = 2.98$
		$_{.95}q_{2,60} = 2.83$	$_{".95"}q_{2,60} = 2.83$

For the Tukey method, the critical value for q is constant for all H_0's in the set.

For the Newman-Keuls method, the largest $\bar{X}_i - \bar{X}_j$ is tested first, hence there are J means being considered and the critical q-value is identical with that for the Tukey. If that is significant, the researcher proceeds to test the second largest mean difference, in which r = J - 1 and the critical q-value is $_{1-\alpha}q_{J-1,f_e}$ which is smaller than when r = J. This procedure is continued until the investigator finds the largest mean differences in the subset being examined to be non-significant at which time he does not continue testing further among the means contained in that particular non-significant subset of means.

The Duncan Multiple Range test is a procedure identical with the Newman-Keuls except that the true α is always greater than the tabled value (except when r = 2). This fluctuation in the true α-value is a feature most consider to be undesirable. For this reason, many authorities never consider Duncan to be the "method of choice."

The F-Distribution Methods

The Scheffe and Planned Orthogonal Contrast (POC) methods are computationally identical for a given H_0, but differ in the critical F-value needed to reject H_0. Both estimate a contrast, ψ, by the expression:

$$\hat{\psi} = C_1\bar{X}_1 + C_2\bar{X}_2 + \ldots + C_J\bar{X}_J$$

The H_o being tested is determined by the values the researcher selects for the C coefficients. Meaningful contrasts require the C's to sum to zero. For example to test $H_o: \mu_1 = \mu_2$, C_1 and C_2 will be 1 and -1. For all pairwise contrasts, the C-values for the two groups will be 1 and -1. (The values for complex contrasts (comparisons involving three or more groups) are not considered in this section). The sum of squares (SS) and the means square for the contrast (since each contrast has one degree of freedom) is:

$$MS_{\psi} = \frac{(C_1\bar{X}_1 + C_2\bar{X}_2 + \ldots + C_J\bar{X}_J)^2}{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \ldots + \frac{C_J^2}{n_J}};$$

or for pairwise contrasts with equal n's:

$$MS_{\psi} = \frac{(\bar{X}_i - \bar{X}_j)^2}{\frac{2}{n}}$$

The F-test is used to test $H_o: \mu_i = \mu_j$, i.e.,

$$F = \frac{MS}{MS_e} \text{ or } \frac{(\bar{X}_i - \bar{X}_j)^2}{\frac{2MS_e}{n}} \text{ or } \frac{(\bar{X}_i - \bar{X}_j)^2}{\frac{2(22)}{11}} = \frac{(\bar{X}_i - \bar{X}_j)^2}{4}$$

(Note: $F = t^2 = 1/2q^2$ for a given comparison)

Is the F-value large enough to be significant, i.e., to reject H_o? The non-directional $H_o: \mu_i = \mu_j$ is rejected at the α-level if the obtained F-ratio exceeds the critical values shown below:

	For Scheffe	For POC
General expression	$(J - 1)\,{}_{1-\alpha}F_{J-1,f_e}$	${}_{1-\alpha}F_{1,f_e}$
Or, in our example with α = .05	$F \geq 5\,{}_{.95}F_{5,60} = 5(2.37) = 11.85$	$F \geq {}_{.95}F_{1,60} = 4.00$

The critical value for the Scheffe test will usually be much larger than the corresponding value for POC. (In this case 11.85 vs. 4.0 or almost 3 times larger for POC). However, POC can test only J - 1 H_0's whereas Scheffe can be used for any number of conceivable H_0's. In addition, like the Dunn test, the POC requires that the H_0's to be tested must be specified prior to the analysis. The J - 1 comparisons must also be orthogonal (i.e., independent). Contrasts will be orthogonal only when the products of the corresponding C's for the two contrasts ψ_a and ψ_b, sum to zero, i.e., $C_{1a}C_{1b} + C_{2a}C_{2b} + \ldots + C_{Ja}C_{Jb} = 0$.

Comparisons Among Methods for our Example

Although the degree of difference between the methods will vary, depending on J and n, the rank order of the magnitude of the differences in means needed to reject $H_0: \mu_i = \mu_j$ is predictable (except when J = 2, when all methods give identical results.) Table 2 gives the magnitude of $\bar{X}_i - \bar{X}_j$ needed to reject $H_0: \mu_i = \mu_j$ (in $S_{\bar{X}_i - \bar{X}_j}$ units) for the various r-values and associated number of H_0's to be tested, c.

Table 2

A Comparison of Mean Differences Needed to Reject H_0 for Pairwise Contrasts for Various Multiple Comparison Methods (in $s_{\bar{X} - \bar{X}}$ units) when J = 6 and n = 11

		r = number of means in subset being examined				
	C	2	3	4	5	6
Multiple-t	15	2.00	2.00	2.00	2.00	2.00
Dunnett	5	2.63	2.63	2.63	2.63	2.63
Dunn	5	2.66	2.66	2.66	2.66	2.66
Dunn	15	3.06	3.06	3.06	3.06	3.06
Tukey	15	2.94	2.94	2.94	2.94	2.94
Newman-Keuls	15	2.00	2.40	2.64	2.81	2.94
Duncan	15	2.00	2.11	2.17	2.22	2.26
Scheffe	15	3.44	3.44	3.44	3.44	3.44
POC	5	2.00	2.00	2.00	2.00	2.00

C=number of pairwise H_0's tested

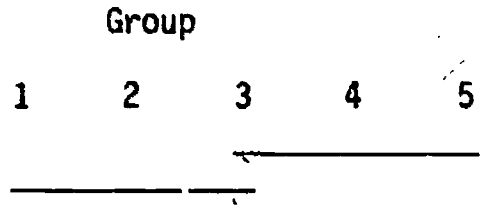
There are other important ways in which these methods differ, one of which is the basis for the α-error rate. Dunn, Dunnett, Tukey and Scheffe use the entire experiment for type I error rate, hence if α = .05, on the course the investigator will make a type I error in only 5% of the experiments he conducts. Newman-Keuls and POC use the individual contrast or H_0's as the basis for α, hence the inbestigator will make a type I error in 5% of the H_0's he tests. Other differences are summarized in Table 3.

[illegible] Summary of Multiple Comparison Methods ([illegible] Prepared by [illegible])

TEST	STATISTICAL BASE	USE	CHIEF ADVANTAGES	CHIEF DISADVANTAGES	[illegible]-BASE	LIMITATIONS, COMMENTS
Multiple t (J . 2)	t	All pairwise comparison	Computational	Many type I errors	[illegible]	Problem of α level
Dunnett	t	Control vs. each other groups (planned)	Most power for those J - 1 contrast	Can't make other interesting comparisons	E	No post-hoc testing -- limited on comparisons.
Dunn or Bonferroni	t	Planned comparisons	Can make all possible simple and /complex comparisons planned	Lower power than for other alternatives if many H_o's tested	E	Use when less than ½ of possible pairwise contrasts are to be tested
Tukey	q	All possible contrasts	Post hoc computation easy	Less powerful than N-K	E	Robust with unequal n's Most useful tecnique with experiment error rate
Newman-Keuls	q	All possible contrasts	Post hoc power for excellent pairwise contrasts	Computational difficulty -- cannot make complex contrasts or set confidence intervals	C	Suggested use as best all-round test for simple contrasts. Confidence intervals not appropriate.
Duncan	q	All possible simple contrasts	Post hoc, good power for pairwise contrasts	Inflated α levels for non-adjacent comparisons	C	Not recommended because of statistical problems with fluctuating α level depending on rank order of means
Planned Orthogonal Contrasts (POC)	F	Planned contrasts which are orthogonal	Greatest power for simple and complex contrast, unequal n's	Limited number of comparisons only J-1 POC	C	Excellent power for few comparisons
Scneffe	F	All possible contrasts (simple and complex) post hoc	All kinds of contrasts, unequal n's	Very little power for testing simple H_o's	E	More powerful than alternative (Except POC) for complex contrasts. Ideal for data snooping.

Depicting Multiple Comparison Results

There are J(J - 1)/2 possible pairwise comparisons. If J is large, for example, 10, then there are 55 hypotheses considered. A parsimonious method of depicting the results is commonly used -- the underscoring procedure. The groups are arranged in order of their means, from low to high. Then each non-significant subgroup is underscored -- any two means underscored by a common line do not differ significantly. Note the example below:

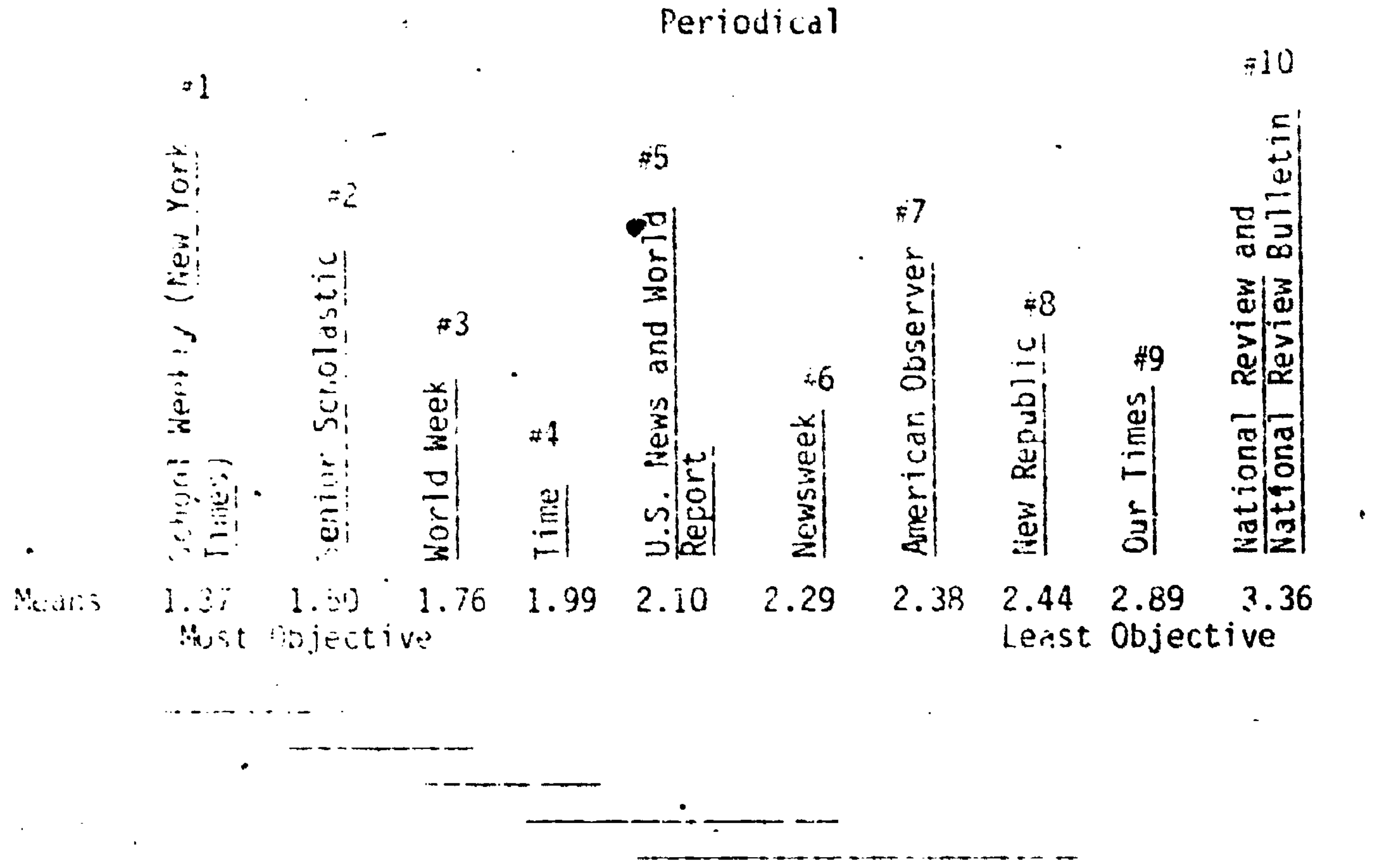


Group 5 differs significantly from groups 1 and 2.
Group 4 differs significantly from groups 1 and 2.
Group 3 does not differ significantly from any group.
Group 2 differs significantly from groups 4 and 5.
Group 1 differs significantly from groups 4 and 5.

Mastery Test: Multiple-Comparison Contrasts

In a study at Cornell University, objectivity ratings given to 10 publications were compared. The dependent variables, whose means are shown in the table below, represent ratings totaled over 13 well defined aspects of objectivity, 4 raters and 3 news events (topics).

Overall Objectivity of All Publications on All Topics (α = .05)



____ 1. According to the results in Table 1, Time (#4) is significantly
a. more objective than publications #'s.
b. less objective than publications #'s.

____ 2. According to the results in Table 1, Our Times is significantly
a. more objective than publication #'s.
b. less objective than publication #'s.

____ 3. The statistical tecniques used to obtain the results shown in the Table were that of
a. planned orthogonal comparisons. b. post-hoc comparisons.

____ 4. Had multiple t-tests been used to compare all the possible pairwise differences, would more "significant" comparisons have resulted"

____ 5. In a given experiment with several groups, which multiple comparison method will require the largest difference between pairs of means in order to reject the null hypothesis; and hence signify the fewest significant differences?
a. Tukey b. Scheffe c. Newman-Keuls

Answers: 1(a): 7-10; 1(b): 1,2; 2(a): 10; 2(b): 1-8; 3: b; 4: yes; 5: b

Multiple Comparisons -- Problem Sets and Notes

Given one fixed ANOVA factor with five treatment levels

SUGGESTION: Use a separate piece of paper to do calculations. Place it over the right-hand side of these sheets, where the answers are given.

	Groups				
	1	2	3	4	5
$\bar{X}$:	47	43	51	54	65
n:	9	9	9	9	9

An ANOVA Summary table is given below

Source of Variation	SS	df	MS	F
Treatments	2520	4	630	6.30
Error	4000	40	100	

$_{.99}F_{4,40} = 3.83$

ANSWERS

1. How many planned orthogonal contrasts (POC) are possible?

4

2. Could you have legitimately inspected the means prior to your selection of the orthogonal contrasts of interest?

No, a priori rationale would no longer apply.

Assume the following definition of the five randomly-assigned groups.

	Abbreviation
1. Control	(C)
2. infrequently tested pupils without feedback	(I,no)
3. frequently tested pupils with negative feedback	(F,-)
4. infrequently tested pupils with positive feedback	(I,+)
5. frequently tested pupils with positive feedback	(F,+)

3. Means for the five groups are given below. Suppose you wished to test $H_0: \mu_3 = \mu_5$. Enter the coefficients for this contrast.

	Group				
	C	I,no	F,-	I,+	F,+
	1	2	3	4	5
$\bar{X}_j$	47	48	51	54	65

$H_0: \mu_3 - \mu_5 = 0$

0, 0, 1, 0, -1 (or 0, 0, -1, 0, 1) i.e., $\psi_1 = \mu_3 - \mu_5 = 0$

4. Suppose you also had good reason to test $H_0: \mu_3 = \mu_4$. Would this be orthogonal with ψ_1?

No

5. Why?

$\Sigma cc \neq 0$; the sum of products of the respective coefficients for the groups must be zero.

6. In addition to $\hat{\psi}_1$, indicate coefficients for the contrast for frequently and infrequently tested groups ($\hat{\psi}_2$).	0,1,-1,1,-1. (or 0,-1,1,-1,1)
7. What is H_0 for $\hat{\psi}_3$ which has 4,-1,-1,-1,-1, as coefficients?	$H_0:\mu_1 = (\mu_2+\mu_3+\mu_4+\mu_5)/4$ or $4\mu_1=\mu_2+\mu_3+\mu_4+\mu_5$
8. Is $\hat{\psi}_2$ orthogonal with $\hat{\psi}_1$?...with $\hat{\psi}_3$?	yes, $\Sigma cc = 0$; yes, $\Sigma cc = 0$
9. Compute $MS_{\hat{\psi}_1}$ $SS_{\hat{\psi}_1} = \dfrac{(\Sigma c\bar{X})^2}{\Sigma \frac{c^2}{n}} = MS_{\hat{\psi}_1}$ since each contrast has df = 1.	$MS_{\hat{\psi}_1} = \dfrac{(-14)^2}{2/9} = 882$
10. Compute F for the contrast $\hat{\psi}_1$.	$F = \dfrac{MS_{\hat{\psi}_1}}{MS_e} = \dfrac{882}{100} = 8.82$
11. For planned orthogonal contrasts (POC) the critical value, $_{.95}F_{1,40}$, in the sample problem is ______	$_{.95}F_{1,40} = 4.08$
12. Is H_0 rejected with $\alpha = .05$.	$_{.95}F = 4.08 < 8.82$ yes, H_0 rejected
13. Would you recommend the POC as the multiple comparison technique in this example.	probably not
14. Why?	Only selected contrasts of interest could be legitimately evaluated.
15. a. Using the Scheffe method (S-method), would you use the identical procedure for obtaining $SS_{\hat{\psi}_1}$ as that for POC?	Yes
b. MS for the contrasts?.	Yes
c. Obtained F for the contrast?	Yes
d. The "critical"value for F?	No, critical F is $(J-1)_{1-\alpha}F_{J-1,f_e}$ for Scheffe, but $_{1-\alpha}F_{1,f_e}$ for POC.

16. For the same H_o using Scheffe method the critical F-value is $(J-1)(_{.95}F_{J-1,f_e})$ or ( )($_{.95}F_{\_\_,40}$) = ( )( ) = 10.44.

$4(_{.95}F_{4,40}) = 4(2.61)$

17. How does this critical value compare with that for orthogonal comparisons, assuming the contrast had been planned?

It is much larger, 10.44 vs. 4.08.

18. Does the S-method require orthogonality?

No, any conceivable contrast is allowable.

19. Does it give a contrast-based type I error rate?

No, an experiment-wise rate

20. For the POC would the critical F-value (4.08) be the same for all four (J - 1) possible comparisons?

Yes

21. Would the critical value for F with the S-method also be constant for all of the possible contrasts.

Yes

22. What is the probability that one or more of the comparisons will yield an F > $(J-1)(_{.95}F_{J-1,f_e})$ when H_o is true?

.05

23. Had the experimenters selected the Tukey method, this distribution theory is no longer based on the F-model but on the

studentized range.

24. Unlike the t which uses $t = \bar{X}_1-\bar{X}_2/s_{\bar{X}_1-\bar{X}_2}$ (or $\bar{X}_1-\bar{X}_2/s_{\bar{X}}\sqrt{2}$, when $n_1 = n_2$) as the critical comparison, Tukey and Newman-Keuls use $q - \bar{X}_1-\bar{X}_2/s_{\bar{X}}$ as the critical ratio on which the distribution theory is based. We should expect, then, that when J = 2, since

$t = \dfrac{\bar{X}_1-\bar{X}_2}{s_{\bar{X}}\sqrt{2}}$, and $q = \dfrac{\bar{X}_1-\bar{X}_2}{s_{\bar{X}}}$, that $t = \dfrac{\rule{2em}{0.4pt}}{\sqrt{\rule{2em}{0.4pt}}}$

$t = \dfrac{q}{\sqrt{2}}$ or $q = \sqrt{2}t$

25. You recall that when J = 2, $F = t^2$, or $t = \sqrt{}$

$t = \sqrt{F}$

26. For the Tukey method the critical value (α = .05) for each comparison would be $_{.95}q_{J,f_e}$ or $_{.95}q_{\_\_,40}$ =

$_{.95}q_{5,40} = 4.04$

27. Since $q_r = \frac{\bar{X}_r - \bar{X}_1}{s_{\bar{X}}}$, a value of $s_{\bar{X}}$ is required that is independent of treatment effects. - Recall the symbol "MS" is another symbol for s^2 and that $s_{\bar{X}}$ and s^2 are related as shown in the equation from elementary statistics:

$s_{\bar{X}} = \frac{s}{\sqrt{n}}$ or $\sqrt{}$

$\sqrt{\frac{MS}{n}}$

28. The n in the above equation is the number of subjects in each group or level, in this case ________.

9

29. Therefore, $s_{\bar{X}} = \sqrt{\frac{(\ )}{(\ )}}$ or

$\sqrt{\frac{100}{9}} = \sqrt{11.1}$ or 3.33

30. Since $s_{\bar{X}}$ and the critical q-value are the same for all Tukey multiple comparisons, the equation can be rearranged so that the minimum significant differences (designated "honest significant differences -- HSD") between a pair of means, HSD = $_{.95}q_{J,f_e}\, s_{\bar{x}}$ = ( ) ( ) = 13.45

(4.04)(3.33)

31. Therefore, in using the Tukey method, every difference between pairs of means greater than 13.45 would be judged significant, and H_o rejected at the ___ level.

.05

The treatment means and the matrix of pairwise differences between treatment means are given below:

$\bar{X}$:	1	2	3	4	5
	47	48	51	54	65

Mean differences

	2	3	4	5
1	1	4	7	18
2		3	7	17
3			3	14
4				11

32. For which differences would H_o be rejected at .05 using the Tukey (HSD) method?

Reject $\mu_1=\mu_5$, $\mu_2=\mu_5$, $\mu_3=\mu_5$

33. Did the S-method reject H_o:$\mu_5=\mu_3$?

No, (cf items 9, 10, 16, F=9.8 < 10.44)

34. The latter computation illustrates the typical relationship between the power of the T and S methods. In comparing pairs of means the _____ method is more efficient and powerful. On the other hand, the S-method is more sensitive in evaluating complex hypotheses, e.g., H_o: $(\mu_1+\mu_2)/2 = (\mu_3+\mu_4+\mu_5)/3$

Tukey

35. The Newman-Keuls (N-K) method, unlike the T and S methods, but like the orthogonal contrasts, bases its type I error rate on the individual comparison rather than on the _________.

experiment

36. The N-K always has J - 1 different critical values for q, or equivalently ______ minimum critical mean differences. Critical q-values are given in tables of the studentized range statistic.

J - 1

37. $q_5(5,40)$ where the means are 5 steps apart = ______ 4.04
$q_4(4,40)$ where the means are 4 steps apart = ______ 3.79
$q_3(3.40)$ where the means are 3 steps apart = ______ 3.44
$q_2(2,40)$ where the means are 2 steps apart = ______ 2.86

38. The minimum differences for r = 5, the extreme-most means, then is identical with that for the ______ method. This is always the case.

Tukey

39. Therefore, for r = 5, minimum mean difference = 13.45.
r = 4, minimum mean difference = (3.79)(   ) = 12.62
r = 3, minimum mean difference = (3.44)(3.33) = 11.46
r = 2, minimum mean difference = (2.86)(3.33) = 9.52

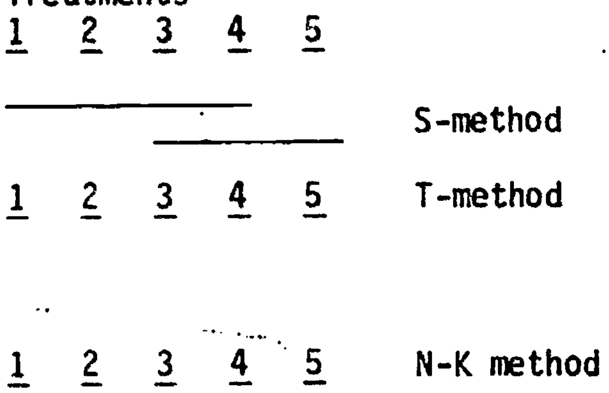
3.33

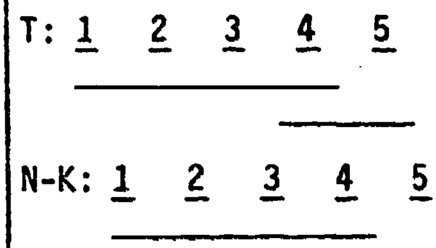
40. Which H_o was rejected for N-K that was not with the T-method? H_o:

H_o: $\mu_4 = \mu_5$

41. Complete the summary figure (any two means not underlined by the same line differ significantly at the .05 level).

Treatments





42. The Dunnett uses the ______as the base for α-error.

43. One group (usually the ______) is compared with each and every other group (or level).

44. In essence the t-ratio is computed, i.e.,

$$t = \frac{\bar{X}_E - \bar{X}_C}{\sqrt{\frac{2MS_e}{n}}} = \frac{\bar{X}_E - \bar{X}_C}{\sqrt{\frac{2(\ \ )}{(\ \ )}}} = \frac{\bar{X}_E - \bar{X}_C}{4.71}$$

45. Since both the critical Dunnett t and $s_{\bar{X}_E - \bar{X}_C}$ are the constant for all comparisons, determining the minimum mean differences will expedite computation, i.e., Min($\bar{X}_E - \bar{X}_C$) = ${}_{.975}t_{J,f_e} s_{\bar{X}_E - \bar{X}_C} = {}_{.975}t(5,40) s_{\bar{X}_E - \bar{X}_C}$ = ( )(4.71) = 12.58

46. How does this compare with the critical mean differences?
a. for the Tukey method?
b. for N-K?

Confidence Intervals (We shall use α = .05)

Planned Contrasts
$\hat{\psi} \pm {}_{.95}F_{1,f_e} \quad MS_e \Sigma \frac{c^2}{n}$ where $\hat{\psi} = \Sigma c\bar{X}$

The nature of the confidence is more apparent when we limit the C.I. to the difference between a pair of means, hence:

$$\hat{\psi} = (1)(\bar{X}_1) + (-1)(\bar{X}_2) \text{ or } \bar{X}_1 - \bar{X}_2$$

Therefore, for planned orthogonal Comparisons between pairs of means, the .95 C.I. is given by

$$\bar{X}_1 - \bar{X}_2 \pm \sqrt{({}_{.95}F_{1,f_e}) MS_e \Sigma \frac{c^2}{n}}$$

experiment

control

$\sqrt{\frac{2MS_e}{n}} = \sqrt{\frac{2(100)}{9}} = \sqrt{22.22} =$
4.71

(An essential difference in the Dunnett and methods "t" is that the critical t-value for the Dunnett considers the fact that there are J groups and J comparisons, not just two.

2.67

smaller
larger than two, smaller than two

Scheffe:

$$\bar{X}_1 - \bar{X}_2 \pm \sqrt{(J - 1)({}_{.95}F_{J-1,f_e})MS_e\Sigma\frac{c^2}{n}}$$

Dunn:

$$\bar{X}_1 - \bar{X}_2 \pm ({}_{.975}t_{c,f_e})\sqrt{2MS_e/n}$$

Tukey:

$$\bar{X}_1 - \bar{X}_2 \pm ({}_{.95}q_{J,f_e})\sqrt{MS_e/n}$$

47. Notice that the confidence interval for the S-method will be larger than that for the orthogonal contrasts to the extent that their respective critical F-value differs:

$\sqrt{(4)\,{}_{.95}F_{4,40}} = \sqrt{(4)(2.61)} =$

$\sqrt{10.44} = 3.23$ is greater than $\sqrt{({}_{.95}F_{1,f_e})}$ or

$\sqrt{4.08} = 2.02$

$4({}_{.95}F_{4,40})$.

48. In the present example 3.23 vs. 2.02 indicates the confidence interval using the S-method is $\left[\frac{3.23}{2.02}\right]$ = 1.6 times greater than the C.I. for a

____

Planned Orthogonal Contrast (POC)

The precision of the estimates can be seen from the relative value for the .95 C.I. for the various multiple comparison approaches in estimating $|\mu_3 - \mu_5|$. (c = number of hypotheses to be tested.)

Orthogonal ($c \leq 4$)	18.06	Dunn (c = 10)	26.56
Tukey	25.54	Dunn (c = 5)	24.22
Scheffe	28.86	Dunnett (c = 4)	23.06

Instructional Exercises

1. If an experiment error rate for the probability of a type I error, α, is desired and hypotheses involve all and only pairs of means, one should select: ____________

Tukey

2. Which of the methods can test all pairs of means and has a <u>contrast</u> error rate? ____________

Newman-Keuls

3. If one were only interested in comparing each $\bar{X}_1$, $\bar{X}_3$, $\bar{X}_4$, and $\bar{X}_5$ with $\bar{X}_C$, he would probably select ____________

Dunnett

4. Which method is most general and places fewest restrictions on the hypotheses that can be tested? ____________

Scheffe

5. In which method will the actual probability of a type I error, α, usually be much larger than the tabled and reported α? ______________

Duncan

6. When there are three or more comparison groups, when will Scheffe necessarily differ from planned orthogonal contrast?

   a. in computing $\hat{\psi}$
   b. in calculating $MS_{\hat{\psi}}$
   c. in computing F
   d. in the appropriate critical F-value
   e. in the coefficients employed for a given contrast

----

d

7. When comparing the extreme-most means:

   a. Tukey will be less powerful than Newman-Keuls
   b. Newman-Keuls will be less powerful than Tukey
   c. Both will be equal in power

----

c

8. In the above situation, will both Tukey and Newman-Keuls be more powerful than Scheffe? ____________

----

yes

9. Will both Tukey and Newman-Keuls be more powerful than Dunn if all pairwise contrasts are to be made? __________

----

yes

10. When J = 3, and $\mu_3 > \mu_2$, the probability of a type II error would be greatest if one employed:

a. Dunnett's technique
b. Newman-Keuls technique
c. Scheffe's technique
d. Tukey's technique

----

c ✓

References

Bailey, D.E. Probability and Statistics. New York: John Wiley & Sons, 1971.

Brownlee, K.A. Statistical Theory and Methodology. New York: John Wiley & Sons, 1965.

Dayton, C.M. The Design of Educational Experiments. New York: McGraw-Hill, 1970.

DuBois, P.H. An Introduction to Psychological Statistics. New York: Harper & Row, 1965.

Duncan, D.B. A Bayesian approach to multiple comparisons. Technometrics, 1965, 7, 171-222.

Dunn, O.J. Multiple comparisons among means. Journal of the American Statistical Association, 1961, 56, 52-64.

Edwards, A.L. Statistical Methods. New York: Holt, Rinehart, & Winston, 1967.

Edwards, A.L. Experimental Design in Psychological Research. New York: Holt, Rinehart, & Winston, 1968.

Edwards, A.L. Statistical Analysis. New York: Holt, Rinehart, & Winston, 1969.

Ferguson, G.A. Statistical Analysis in Psychological Education. New York: McGraw-Hill, 1971.

Fryer, H.C. Concepts and Methods of Experimental Statistics. Boston: Allyn and Bacon, 1966.

Glass, G.V & Stanley, J.C. Statistical Methods in Education and Psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Guilford, J.P. Fundamental Statistics in Psychology and Education (3rd ed.). New York: McGraw-Hill, 1965.

Hays, W.L. Statistics for Psychologists. New York: Holt, Rinehart, & Winston, 1963.

Hays, W.L. & Winkler, R.L. Statistics: Probability, Inference and Decision. New York: Holt, Rinehart, & Winston, 1971.

Hopkins, K.D. An empirical analysis of the efficacy of the WISC in the diagnosis of organicity in children of normal intelligence. Journal of Genetic Psychology, 1964, 105, 163-172.

Hopkins, K.D. Preventing the number one misinterpretation of behavioral research or how to increase statistical power. Journal of Special Education, 1972, (in press).

Hopkins, K.D. & Chadbourn, R.A. A schema for proper utilization of multiple comparisons in research and a case study. American Educational Research Journal, 1967, 4, 407-412.

Kirk, R.E. Experimental Design: Procedures for the Behavioral Sciences. Belmont, Calif.: Brooks-Cole Publishing Co., 1968.

Li, C.C. Introduction to Experimental Statistics. New York: McGraw-Hill, 1964.

Marascuilo, L.A. Large sample multiple comparisons. Psychological Bulletin, 1966, 65, 280-290.

Marascuilo, L.A. Statistical Methods for Behavior Science Research. New York: McGraw-Hill, 1971.

McNemar, Q. Psychological Statistics. New York: Wiley, 1969.

Miller, R.G. Simultaneous Statistical Inference. New York: McGraw-Hill, 1966.

Myers, J.L. Fundamentals of Experimental Design. Boston: Allyn and Bacon, 1972.

Ostle, B. Statistics in Research. Ames, Iowa: Iowa State University Press, 1963.

Petrinovich, L.A. & Hardyck, C.D. Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. Psychological Bulletin, 1969, 71, 43-54.

Roscoe, J.T. Fundamental Research Statistics. New York: Holt, Rinehart, & Winston, 1969.

Scheffé, H. The Analysis of Variance. New York: John Wiley & Sons, 1959.

Snedecor, G.W. & Cochran, W.G. Statistical Methods. Ames, Iowa: Iowa State University Press, 1967.

Sparks, J.N. Expository notes on the problem of making multiple comparisons in a completely randomized design. Journal of Experimental Education, 1963, 31, 342-349.

Steel, R.G.D. & Torrie, J.H. Principles and Procedures of Statistics. New York: McGraw-Hill, 1960.

Tringo, J.L. The hierarchy of preference toward disability groups. Journal of Special Education, 1970, 4, 295-306.

Walker, H.M. & Lev, J. Elementary Statistical Methods. New York: Holt, Rinehart, & Winston, 1969.

Wilson, E.D. A comparison of the effects of deafness simulation and observation upon attitudes, anxiety, and behavior manifested toward the deaf. Journal of Special Education, 1971, 5, 343-349.

Winer, B.J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1971.