Multiple Comparisons of Means: A Practical Guide

Gary M. Ingersoll
United Arab Emirates University

Abstract: Analysis of variance (ANOVA) is a powerful set of statistical procedures that allows a researcher to compare relative differences among treatment conditions or types of individuals. In this review a limited set of post-ANOVA multiple comparison techniques are described and a set of general guidelines are provided that will accommodate most situations.
Keywords: Analysis of variance, multiple comparisons

Analysis of variance (ANOVA) is a powerful set of statistical procedures that allows a researcher to compare relative differences among treatment conditions or types of individuals. It is a common misconception, however, for people to say that ANOVA tests for differences among group means. Actually, the test of significance (the F-ratio) is a ratio of the observed variability of group means to the expected random variability of group means based on the theory of the sampling distribution of means. A statistically significant F-ratio simply says that the observed variability was greater than chance variability, hence its name. Once overall statistically significant differences have been demonstrated to exist among treatment means using an overall F-ratio, however, the investigator is left with the question: "Which means are statistically different and which are not?" A significant F-ratio fails to tell where differences are found. It is to the process of identification of determining how means differ that this paper is directed.

Specifically, my purpose here is to describe selected elements of multiple comparison techniques and to offer some general guidelines about a subset of techniques that will accommodate most situations; it will focus on different analytic settings and offer heuristic guidelines about which multiple comparison technique to use in each setting. It is not the intent to review a broad base of techniques. An interested reader is referred to resources such as Hochberg and Tamhane (1987), Hsu (1966) or Toothaker (1993) for that purpose.

Professor Paul Games (1971) once lamented “The area of multiple comparisons is one of the more confusing areas of statistics, and is one that

receives a widely differing set of recommendations from many applied statistics tests in the behavioral sciences (p.531).” More recently, Sato (1996) noted that the wide variety of possible multiple comparison techniques lead to confusion and frustration. One might argue that the problem remains although not quite as badly as at the time of Games’s ruminations. Twenty-five years after Games, Hancock and Klockars (1996) offered an updated view but one that is not necessarily consonant with Kirk (1995), Winer (1971, Winer, Brown, & Michaels, 1991) or some others. The conceptual model offered herein follows the procedures outlined by Games (1971).

As in the overall, omnibus analysis, we need to strike an appropriate balance between our risk of Type I and Type II error. Given a statistically significant F-ratio, the investigator is justified in proceeding with further analyses of treatment means. Typically this analysis is guided by a set of preplanned (or *a priori)* comparisons. Sometimes, however, the investigator does not have a preplanned set of comparisons. In the latter case, exploration of differences among means should proceed with a more conservative test. The reason for the increased caution is that since the investigator is unaware of where treatment means should differ, any exploration may capitalize on chance differences.

Consider data from the following study (Ingersoll, Orr, Vance, & Golden, 1992). The study was directed at variables that contribute to effective self management of Type 1 (Insulin-Dependent) Diabetes among adolescents. Better self-management results in better metabolic control. Adolescents with Type 1 Diabetes are notoriously in poor metabolic control increasing the risks of long-term complications of the disease. Self-management requires a complex balance of insulin injections, diet, and exercise. Too little insulin and exercise in the presence of too much food results in elevated blood sugar levels (hyperglycemia). Too much insulin in the face of lowered exercise and diet results in lowered blood sugar levels (hypoglycemia). Both conditions can be dangerous, even life threatening.

Data from that study were subjected to a factorial ANOVA focusing on two independent variables. First, there is a well-established literature that indicates that adolescent girls are routinely in poorer metabolic control than adolescent boys. Second, since the self-management involves a complex regimen, we included a measure of cognitive social development known as conceptual level (CL). The adolescents were divided into three groups: Low CL learners who are conceptually rigid and simple, Moderate CL learners who are somewhat more flexible but still need well-structured environments, and High CL learners who are flexible and adaptable. The resulting data are thus conceived as a 2 by 3 factorial analysis. The

dependent variable was a measure of metabolic control called glycosylated hemoglobin which gives an estimate of average quality of metabolic control over a period of about 60 days. A lower value indicates better average metabolic control. Table 1 presents the ANOVA summary table for the data and Table 2 the cell and marginal means. Analysis of variance revealed statistically significant differences for both main effects and no interaction. That is, the test for metabolic control by gender yielded $F(1,118)=9.07$, $p=.003$ and the test by CL yielded $F(2,118)=8.39$, $p<.001$.

Table 1

ANOVA Summary Table SPSS Output Using General Linear Model

Tests of Between-Subjects Effects

Dependent Variable:hba1

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	129.013[a]	5	25.803	6.814	.000
Intercept	12424.595	1	12424.595	3281.206	.000
Sex	34.356	1	34.356	9.073	.003
Cog	63.546	2	31.773	8.391	.000
sex * cog	1.526	2	.763	.202	.818
Error	446.818	118	3.787		
Total	14128.439	124			
Corrected Total	575.831	123			

a. R Squared = .224 (Adjusted R Squared = .191)

The question at hand is “Do the mean levels of metabolic control differ among the three groups?” Note, since there are only two groups in the gender dimension no further tests are needed. Females were in less well controlled metabolic status than males. We need only to pursue differences among the CL groups.

Table 2
Mean Glycosylated Hemoglobin Levels of Males and Females at Three Levels of Conceptual Development

	Conceptual Level			
	Low	Mod	High	All
n_j	34	52	38	124
Sex				
Male	10.95	9.57	9.30	9.74
Female	12.22	10.84	10.07	11.15
All	11.77	10.23	9.58	10.45

Four approaches will be reviewed using both computational and computer output approaches: the Fisher LSD, the Tukey HSD, the Dunn, and the Scheffé approaches.

Pairwise Multiple Comparisons

One logical alternative to post-ANOVA comparisons of means would be to modify our original t-test by using the mean square within as a more appropriate estimate of variability of sample means. We could then compare all pairs of treatment means $\overline{Y}_{.i} - \overline{Y}_{.j}$ testing all possible pairwise differences against the null hypotheses of no difference.

At this point, one might wonder "Why not just do a series of simple t-test?" The answer relates to what is called experiment-wise error rate. Experiment wise error rates refer to the number of comparisons within an experiment in which we expect to find *at least one* Type I error. If we start with a risk of type 1 error (α) set at .05, repeated t-tests degrade that error. With 3 or more groups multiple comparisons using a simple t is no longer an acceptable procedure (Kirk, 1968). What we want is a technique that maintains a constant risk of type 1 error across comparisons.

Fisher LSD

The Fisher Least Significant Difference (LSD) approach is based on an analog to the simple t-test using the mean square for the error term (MS_{error}) as an unbiased estimate of error variance. In so doing we arrive at the formula:

$$t_{df=MSe} = \frac{\overline{Y}_{.i} - \overline{Y}_{.j}}{\sqrt{MS_{error}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

We would then compare the resulting t-test to the critical value $t_{df=error}^{1-\alpha/2}$. If the computed ***t*** exceeds the critical $t_{df=error}^{1-\alpha/2}$ we reject the null hypothesis. In this instance, looking at Table 1, we find that $MS_{error} = 3.787$ and the degree of freedom for the error term is 118. The two-tail critical value for a t-test with 118 degrees of freedom is 1.98.[i] Comparing the means of Low CL and High CL adolescents can be computed as

$$t_{df=MSe} = \frac{\overline{Y}_{.i} - \overline{Y}_{.j}}{\sqrt{MS_{error}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} = \frac{11.77 - 9.58}{\sqrt{3.787(\frac{1}{34} + \frac{1}{38})}} = \frac{2.19}{\sqrt{3.787(.029 + .026)}} = 4.77$$

Since the observed value exceeds the critical value we reject the null hypothesis of no difference. In the case of pairwise comparisons, SPSS provides the option (under General Linear Model, Univariate, Post Hoc) for a test of pairwise differences. Table 3 provides a SPSS summary of the tests of pairwise differences using the Fisher LSD.

Interpreting the SPSS output is quite direct. The program provides the pairwise difference for all combinations and notes whether the difference is statistically significant at the .05 level with a "*". Note that there is redundancy in the table. The comparison of group 1 versus group 3 is the

same as group 3 versus group 1. Looking at the indicators of statistical significance, we find that group 3 differs for both group 1 and group 2 but that groups 2 and 3 do not differ.

The output also provides the standard error for the pairwise comparison and you will note that the standard error for comparison 1 versus 3 matches the computation above. The output also includes the "exact" probabilities for each comparison. When the probability is less than .001, it simply prints .000.

Table 3

Multiple Pairwise Comparison of Means SPSS Output with LSD

Multiple Comparisons

hba1

LSD

(I) cog	(J) cog	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00	2.00	1.5401*	.42917	.000	.6902	2.3900
	3.00	2.1883*	.45937	.000	1.2786	3.0979
2.00	1.00	-1.5401*	.42917	.000	-2.3900	-.6902
	3.00	.6481	.41529	.121	-.1743	1.4705
3.00	1.00	-2.1883*	.45937	.000	-3.0979	-1.2786
	2.00	-.6481	.41529	.121	-1.4705	.1743

Based on observed means.

The error term is Mean Square (Error) = 3.787.

*. The mean difference is significant at the .05 level.

Perhaps the more useful element that should not be ignored is the 95% confidence interval. The 95% confidence interval is computed by the observed difference plus or minus the critical value times the standard error.

$$CI_{1-\alpha} = \left(\overline{Y}_i - \overline{Y}_j\right) \pm t_{1-\alpha/2} * S_{error}$$

That is, our observed difference of 2.19 is an estimate. It contains error. We are however, willing to conclude that 95 percent of the time, the real difference between groups 1 and 3 falls somewhere between 1.28 and 3.10. Further, the confidence interval does not overlap with the null hypothesis of no difference.

The principal problem with the Fisher LSD approach is that it retains too high an experiment-wise risk of Type 1 error, especially in the context of pairwise comparisons. Most writers discourage its use. Thus, while conveniently available, the Fisher LSD probably should be avoided.

Tukey's HSD

The Tukey Honestly Significant Difference (HSD) procedure is similar to the LSD procedure but uses a modified error term and compensates for the increased experiment wise error rate that results from comparing all pairwise means through a modification of the critical comparison value. The Tukey HSD is a *post hoc* procedure. That is, there are no preplanned or *a priori* hypotheses of expected differences. As a *post hoc* comparison technique, all tests are two-tailed.

Computing the Q statistic for Tukey's HSD requires a modest alteration of the LSD formula. The standard error of the difference is the square root of the mean square error divided by the cell size.

$$Q_{k,dfe} = \frac{\overline{Y}_i - \overline{Y}_j}{\sqrt{MS_{error}/n}}$$

The original Tukey's HSD thus required equal cell sizes. Multiple alternative approaches are available to compensate for unequal cell sizes but a simple alternative is to compute a harmonic mean of cell samples. That is:

$$Q_{k,dfe} = \frac{\overline{Y}_i - \overline{Y}_j}{\sqrt{MS_{error}/n_{harmonic}}}$$

where

$$\bar{n}_{harmonic} = \frac{j}{\sum \frac{1}{n_j}} = \frac{3}{\frac{1}{34} + \frac{1}{52} + \frac{1}{38}} = 40.02$$

Thus, continuing to pursue our difference between groups 1 and 3:

$$Q_{k,dfe} = \frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{MS_{error} / n_{harmonic}}} = \frac{11.77 - 9.58}{\sqrt{\frac{3.787}{40.02}}} = 7.11$$

The comparison statistic **Q** is specified as a function of the degrees of freedom for the error term (df_e) and the number of means being compared (**j**, the number of groups). In our sample case, we have 3 group means and 118 degrees of freedom for our error term. The critical value for 118 is not available but we can use either the critical value for 3 groups and df=60 or 3 groups and df=120. The critical values of ${}_{1-\alpha}\mathbf{Q}_{k,dfe}$ are ${}_{.05}\mathbf{Q}_{3,120} = 3.36$ and ${}_{.01}\mathbf{Q}_{3,120} = 4.20$.

In reviewing a table of critical values for Tukey's HSD, note that for each row (df_e), the weighted comparison statistic becomes increasingly conservative as the number of treatment means increases. That is, given a constant number of degrees of freedom for our error term, as the number of comparisons increases, the critical value ${}_{1-\alpha}\mathbf{Q}_{k,dfe}$ for the Tukey test also increases. This reflects first the increased number of possible comparisons, but also the decreased number of observations per comparison. That is, in the current case, the number of observations per mean ($\mathbf{n_j}$) is about 40. If we were to generate comparisons for 6 means with an error term with 120 degrees of freedom, the number of observations per mean would be 20. As we decrease the number of observations per treatment mean, its stability is diminished. Likewise, if we increase our number of treatment means from 3 to 6, we increase the number of possible pairwise comparisons from 3 to 15. Both factors lead us to be more cautious in our evaluation of pairwise differences. We want to increase our protection from a risk of Type I error.

Table 4
Multiple Pairwise Comparison of Means SPSS Output with HSD

Multiple Comparisons

hba1

Tukey HSD

(I) cog	(J) cog	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00	2.00	1.5401*	.42917	.001	.5214	2.5588
	3.00	2.1883*	.45937	.000	1.0979	3.2786
2.00	1.00	-1.5401*	.42917	.001	-2.5588	-.5214
	3.00	.6481	.41529	.267	-.3376	1.6339
3.00	1.00	-2.1883*	.45937	.000	-3.2786	-1.0979
	2.00	-.6481	.41529	.267	-1.6339	.3376

Based on observed means.

The error term is Mean Square (Error) = 3.787.

*. The mean difference is significant at the .05 level.

As with the LSD, SPSS provides the option for the Tukey test of pairwise differences. Table 4 provides a summary of the tests of pairwise differences using the Tukey HSD. Interpreting the SPSS output is comparable to our earlier attention to the LSD. The program provides the pairwise difference for all combinations and notes whether the difference is statistically significant at the .05 level with a "*". Again, there is redundancy in the table. Looking at the indicators of statistical significance, we find that group 3 differs for both group 1 and group 2 but that groups 2 and 3 do not differ. However, where the exact probability of the 2 versus 3 comparison in the case of the LSD comparison was .121, it was .267 for the HSD comparison, indicating a more conservative view.

Again the more useful element that should not be ignored is the 95% confidence interval. The 95% confidence interval is computed by the

observed difference plus or minus the critical value times the standard error.

$$CI_{1-\alpha} = \left(\overline{Y}_i - \overline{Y}_j\right) \pm Q_{k,dfe} * S_{error}$$

Again, our observed difference of 2.19 is an estimate. It contains error. We are however, willing to conclude that 95 percent of the time, the real difference between groups 1 and 3 falls somewhere between 1.10 and 3.28. The confidence interval is now wider than the LSD procedure.

The use of all unplanned pairwise comparisons is a statistically weak, inferential procedure. In using this technique we really are unclear about the nature of how differences among treatment means should be distributed in a meaningful fashion. We are saying "We'll take what we can get." We must be cautious in our interpretations of pairwise differences because they are evaluated in the context of the theoretical comparison of all possible questions. The Tukey procedure is most appropriately applied when an investigator wishes to make all pairwise comparisons. By definition, this would be a theoretically weak approach. Hence, the HSD is more conservative. If we have preplanned comparisons the HSD increases our risk of Type 2 error.

Weighted Linear Contrasts

In more complex theoretical models, rather than rely on all pairwise comparisons, we prefer to use a finite set of "weighted" comparisons that are linked to an existing empirical context. These weighted contrasts may be either preplanned (*a priori*) or unplanned (*post hoc*.) The test statistic for the weighted contrast will be the same in both cases but the comparison value will differ for preplanned versus unplanned comparisons.

The test statistic is an extension of the logic of the traditional t-test (Games, 1971.) It is composed of a weighted linear contrast and a standard error of the contrast. These contrasts are linear since they are composed of additive, weighted combinations of all the treatment means are denoted by ψ_i. Each linear contrast is translatable into a null hypothesis $H_0: \psi_i = 0$. Theoretically, the hypothesized linear contrast assigns weights $\left(w_i\right)$ to population means for each experimental condition $\left(\mu_i\right)$ but in practice we estimate the population means by the sample means for each condition $\left(\overline{Y}_i\right)$. Thus $\psi_i = w_1\mu_1 + w_2\mu_2 + w_3\mu_3 + \ldots + w_k\mu_k$ is tested by way of

$\hat{\psi}_i = w_1\overline{Y}_1 + w_2\overline{Y}_2 + w_3\overline{Y}_3 + \ldots + w_k\overline{Y}_k$ which is an estimate. For any valid weighted linear contrast, the sum of the linear weight is 0, i.e., $\sum w_j = 0$.

The standard error of the weighted contrast is $\sqrt{MS_{error}\sum\frac{w_j^2}{n_j}}$. That is, it is the square root of the product of the mean square error and the sum of the squared weights divided by their respective sample sizes. The pairwise comparison is a special case of the weighted linear contrast in which the weights are +1 and -1 which when squared are both 1. The t-test is thus:

$$t_\psi = \frac{\psi_i}{S_\psi} = \frac{\psi_i}{\sqrt{ms_{error}\sum\frac{w_j^2}{n_j}}}$$

Consider the data at hand. In the original analysis, the following hypotheses could be put forward:

1. High CL adolescents with diabetes are in better metabolic control than the remaining adolescents with diabetes.
2. Moderate CL adolescents with diabetes are in better metabolic control than Low CL adolescents with diabetes.

   These two hypotheses can be depicted by the following linear contrasts:

$$H_o: \psi_1 = \left(+\frac{1}{2}\right)\mu_1 + \left(+\frac{1}{2}\right)\mu_2 + (-1)\mu_3$$

$$H_o: \psi_2 = (+1)\mu_1 + (-1)\mu_2 + (0)\mu_3$$

Computing the respective t-tests for the two contrasts we find:

$$t_1 = \frac{\tilde{\psi}_1}{\sqrt{MS_{error}\sum\frac{w_j^2}{n_j}}} = \frac{\left(+\frac{1}{2}\right)\overline{Y}_1 + \left(+\frac{1}{2}\right)\overline{Y}_2 + (-1)\overline{Y}_3}{\sqrt{MS_{error}\sum\frac{w_j^2}{n_j}}}$$

$$t_1 = \frac{\frac{11.77+10.23}{2} - 9.58}{\sqrt{3.787\left(\frac{+.5^2}{34} + \frac{+.5^2}{52} + \frac{-1^2}{38}\right)}} = \frac{1.42}{\sqrt{3.787*.038}}$$

$$t_1 = 3.27$$

$$t_2 = \frac{\tilde{\psi}_2}{\sqrt{MS_{error}\sum\frac{w_j^2}{n_j}}} = \frac{(+1)\overline{Y}_1 + (-1)\overline{Y}_2 + (0)\overline{Y}_3}{\sqrt{MS_{error}\sum\frac{w_j^2}{n_j}}}$$

$$t_2 = \frac{11.77 - 10.25}{\sqrt{3.787\left(\frac{+1^2}{34} + \frac{+1^2}{52}\right)}} = \frac{.65}{\sqrt{3.787*.049}}$$

$$t_2 = 1.51$$

The question now becomes, "Are these contrasts statistically significant?" The answer is tied to whether the weighted contrasts are preplanned or unplanned and exploratory. Both the Dunn procedure and the Scheffé procedure apply the same test statistic. They differ in their critical values.

Dunn

When we have contrasts that are preplanned (hypothesized) and are restricted to those that have theoretical meaning we gain statistical power. That is, we reduce our risk of Type 2 error and the comparison techniques that we use to reflect that gain. The Dunn t-test (or sometimes the Dunn-Sidak or Bonferroni) uses the same test statistic as the Scheffé, however, the comparison statistic is based on the premises that 1) comparisons are preplanned and 2) the critical value of the test statistic is directly related to

the number of planned comparisons and the degrees of freedom of the error term.

The critical value of the Dunn statistic is found by accessing a table of values and is tied to the number of comparisons (in this instance 2) and the degrees of freedom for the mean square error (118). Since your hypotheses are preplanned, they should be directional. Accessing the table we find that the critical value for Dunn α=.05 is 2.43 and for α=.01 is 2.99. Thus, we reject the null hypothesis for hypothesis 1 but retain hypothesis 2.

Although some might argue differently, there are times when the most desirable and informative approach to data analysis following ANOVA is a preplanned analysis of pairwise comparisons. With a small number of groups, this is manageable and appropriate with the Dunn procedure. In the present example 3 groups result in 3 pairwise comparisons. But, four groups yields 6 and five groups yields 10. Very quickly, any advantage of preplanning is lost.

Confidence Interval

As noted before, the 95% confidence interval is computed by the observed difference plus or minus the critical value times the standard error. Thus

$$P\left[\tilde{\psi}_{obs} - 1-\alpha/2\, Dunn_{(dfe,c)}\ S_{\psi} \le \psi_j \le \tilde{\psi}_{obs} + 1-\alpha/2\, Dunn_{(dfe,c)}\ S_{\psi}\right] = 1-\alpha$$

$$P\left[\hat{\psi}_1 - 2.43\,(0.382) \le \psi_1 \le \hat{\psi}_1 + 2.43\,(0.382)\right] = 1-.05$$

$$P\left[1.42 - .93 \le \psi_1 \le 1.42 + .93\right] = .95$$

$$P\left[0.49 \le \psi_1 \le 2.35\right] = .95$$

Scheffé

As several writers (Games, 1971; Huck, 2000; Kirk, 1995; Hancock & Klockars, 1996) indicate, the Scheffe' technique is the most general, it is also the most constricting. It is preferable for exploratory unplanned *post hoc* weighted linear contrasts. The Scheffé critical value is defined as $t_{\Psi} \ge \sqrt{(a-1)^* F_{(a-1,dfe)}}$ where a is the number of groups and $F_{(a-1)(dfe)}$ is

the critical value for an ANOVA test. The critical value for $F_{2,118}$ is 3.087. Thus

$$t_{\psi} \geq \sqrt{(a-1)^{*} F_{(a-1,dfe)}}$$

$$t_{\psi} \geq \sqrt{(3-1)(3.087} = 2.48$$

Discussion

In his classic volume Fisher (1935) advised that if no significant F-ratio was found in the omnibus F-test, no further action was justified. If a significant ratio was found then the researcher was justified in further exploration. There has been some debate about the first proposition (Davis & Gaito, 1984) especially in the context of the debate on the arbitrariness of the null hypothesis testing model. If a researcher has a priori hypotheses, follow up comparisons may be justified, even in the absence of a significant F-ratio.

Two cautions are in order. First, as Tukey (1991) warns the presentation of an "exact" statistic for a comparison may be misleading. Any treatment mean and thus the difference between treatment means is an estimate and contains error. A confidence interval based on the multiple comparison statistic is preferred. Second, in accord with the current dissatisfaction with rigid adherence to the null hypothesis statistical testing model, statistical significance may not equate with practical significance. A measure of effect size should accompany the test. See Nakagawa and Cuthill (2007) for a useful review of this topic.

The average reader encountering the literature on multiple comparison procedures is apt to throw up his or her arms in frustration. In one review, Miller (1977) cited 255 papers on alternative multiple comparison procedures in an 11 year period, many quite arcane. I agree with Games (1978) that the proliferation of alternative multiple comparison processes is not productive. As Games (1971, 1978) notes, the various methods are all reducible to either a version of Fisher's t or Sheffé's F.

In thinking about multiple comparison procedures, one is always engaged in a trade off of risks of Type 1 and Type 2 error (Cribbie, 2003; Davis & Gaito, 1984; Games, 1971; Sato, 1996). In the context of a priori planned comparisons, we use a more liberal standard that moderately increases risk of Type 1 error but radically diminishes Type 2 error. Conversely, when exploring unplanned post-hoc comparisons, we use a more judicious and conservative approach reducing our risk of Type 1 error at the cost of increasing our risk of Type 2 error. Of the techniques, the Scheffe technique is the most conservative

and the Fisher most liberal.

Finally, to review the decision about which multiple comparison technique to apply relates to a) whether the comparisons are pairwise or weighted, and b) whether the contrasts are preplanned (*a priori*) or unplanned (*post hoc*) and exploratory. In general the following guidelines are in order:

1. Pairwise comparisons are most often post-hoc and thus the Tukey procedure is most appropriate.
2. In the instance of pre-planned pairwise comparisons (an expected ordered outcome such as in the example case) the Dunn process is most appropriate.
3. Weighted linear contrasts that are preplanned should be tested using the Dunn process.
4. Weighted linear contrasts that are unplanned should be tested using the Scheffé process.

[i] Tables of critical values for all the tests noted here are available in most intermediate texts of statistics or are readily available on-line.

References

Cribbie, R. A. (2003). Pairwise multiple comparisons: New yardstick, new results. *Journal of Experimental Education, 71*(3), 251-265.

Davis, C., & Gaito, J. (1984). Multiple comparison procedures within experimental research. *Canadian Psychology, 25*(1), l-13.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, UK: Oliver & Boyd.

Gaito, J., & Nobrega, J. A. (1981). A note on multiple comparisons as an ANOVA problem. *Bulletin of the Psychonomic Society, 17*(3), 169-170.

Games, P. A. (1971). Multiple comparison of means. *American Educational Research Journal, 8*(3), 531-565.

Games, P. A. (1987). A three factor model encompassing many possible statistical tests on independent groups. *Psychological Bulletin, 85*(1), 168-182.

Hancock, G. R., & Klockars, A. J. (1996). The quest for alpha: Developments in multiple comparison procedures in the quarter century since Games (1971). *Review of Educational Research, 66*(3), 269-306.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York, NY: John Wiley & Sons.

Hsu, J. C. (1966). *Multiple comparisons: Theory and methods*. Boca Raton, FL: Chapman & Hall/CRC.

Ingersoll, G. M., Orr, D. P., Vance, M. D. & Golden, M. P. (1992). Cognitive maturity, stressful events and metabolic control among diabetic adolescents. In E. J. Susman, L. V. Fegans & W. J. Ray (Eds) *Emotion, Cognition, Health and Development in Children and Adolescents*. Hillsdale, NJ: Lawrence Erlbaum.

Kirk, R. (1995). *Experimental design: Procedures for the behavioral sciences.* Pacific Grove, CA: Brooks/Cole.

Miller, R. G. (1977). Developments in multiple comparisons: 1966-1976. *Journal of the American Statistical Association*, *72*(360), 779-788.

Nakagawa, S., & Cuthill, I. C. (2007), Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews, 83*(4), 591-605. Accessed 8 March 2010.

Sato, T. (1996). Type I and type II errors in multiple comparisons. *Journal of Psychology, 130*(3), 293-302.

Toothaker, L. E. (1993). *Multiple comparison procedures.* Newbury Park, CA: Sage.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*(1), 100-116. Accessed 8 March 2010.

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

Winer, B. J., Brown, D. R., & Michaels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw- Hill.