How Should We Teach Follow-Up Tests After Significant Interaction in Factorial Analysis of Variance?

T. C. Oshima
Georgia State University

Frances McCarty
Emory University


Key Words: Factorial ANOVA, Interaction, Follow-Up Tests, Textbooks

Abstract: Although factorial analysis of variance (ANOVA) is a widely used statistical technique in social sciences, most textbooks do not sufficiently explain what to do after a significant interaction effect. In this paper, sources of problems are discussed from the technical view as well as from the students' point of view. A survey of textbooks was conducted to confirm our speculation. A student handout (available on the web) was developed so that students whose textbooks are lacking in explaining how to do the follow-up tests in factorial ANOVA can conduct his or her research using SPSS.

1. Introduction

Although most of the introductory and intermediate statistics textbooks include the topic of factorial analysis of variance, the treatment of how to perform follow-up tests is rather cursory. The information on how to proceed after a significant interaction is either lacking or confusing. Huck (2000) suggests three strategies to gain insight into a statistically significant interaction:

Strategy 1: To investigate the graph of cell means.
Strategy 2: To perform a statistical comparison of cell means.
Strategy 3: To conduct tests of simple main effects.

To facilitate the illustration let us use an example here. Suppose two factors are type of drug (3 levels) and gender (2 levels). In the first strategy, one is to look at the graph of cell means and evaluate the non-parallel lines. The significant F in the interaction suggests that these lines are not likely to come from parallel lines in the population. The problem with this approach is that students often conclude that cell means on each line are “different”. They may say “for males the mean for Drug 1 is higher than the means for Drugs 2 and 3, and for females the mean for Drug 3 is higher than the other two means, thus suggesting Drug 1 is the best treatment for males and Drug 3 is the best treatment for females”. However, the difference of cell means here is simply descriptive. In other words, the difference may not be statistically significant. Therefore, although this strategy is useful and perhaps a necessary step to investigate the nature of an interaction, it is not a complete strategy.
In Strategy 2, after the significant interaction, all cell means are statistically compared simultaneously perhaps using a follow-up test such as the Scheffé test. This strategy, referred as all pairwise comparison (APC) hereafter, is straightforward and easy to explain to students. Also, explaining how to do this on the computer (using SPSS for example) is fairly simple. However, this strategy is very conservative and produces numerous comparisons that are not of any concern for the researcher. In the example above, it would be useful to compare, say, Drug 1 vs. Drug 2 for males. However, comparing, say, Drug 1 for males and Drug 2 for females seems rather non-important, because the researcher would want to find out which drug to use given the gender of the patients.

Strategy 3, referred to as simple-main effects (SME) analysis hereafter, appears to answer the researcher's substantive question best, but it is the most difficult to teach to students. In this strategy, after the significant interaction, the data are split for each level of one factor and one-way ANOVAs are conducted. In the example above, a one-way ANOVA is conducted for males to find out which drug works the best for males, and another one-way ANOVA is conducted to find out which drug works the best for females. Like any other one-way ANOVA with more than two levels, after the significant F, a post-hoc test such as the Tukey test or the Scheffé test is conducted to find out which pair (or pairs) of means is (are) statistically different.

Although this process makes sense in terms the researcher's inquiry, there are statistical problems associated with Type 1 error ($\alpha$) and also the power. In factorial ANOVA, multiple F tests are already conducted (e.g., Test of main effect A, Test of main effect B, and Test of A x B) and the inflation of Type 1 error rate is a concern. (See Kromrey and Dickinson, 1995 for a discussion of this problem and how to alleviate the problem.) If a series of SME analyses are conducted, this inflation problem continues to exist at the follow-up stage. Another concern is the use of Mean Square Error (MSE) in calculating the F for the simple-main effects. Although it may seem more advisable to use MSE from the factorial ANOVA as opposed to MSE from the follow-up one-way ANOVA, common computer programs such as SPSS automatically use the MSE from the one-way ANOVA. Therefore, students need to hand-calculate F for the follow-up ANOVA if she or he wants to use MSE from the factorial ANOVA.

We have considered all three approaches in our teaching. We find that students have a hard time understanding and conducting Strategy 3. We also found that most of the textbooks we looked at have a very cursory treatment of this problem. Some textbooks stop the explanation after presenting an ANOVA table in which a significant interaction is found. When students conduct their research (e.g., dissertation), they cannot and should not stop at the significant interaction. They need to come up with a recommendation such as which method works the best given certain attributes (often the second factor).

The purpose of this paper is to (1) present the existing problems, (2) survey the major introductory and intermediate statistics textbooks concerning how they deal with these problems, and (3) offer a handout that will help students better understand the follow-up procedures in the factorial ANOVA which are often not explained sufficiently in textbooks.

2. Existing Problems
The problems described above stem from two sources. One source is the technical problems which continue to be debated among statisticians. The other source is on the side of students faced with the factorial designs. We will first discuss the first source.

### 2.1 First Source: Technical Aspects

Although factorial ANOVA in terms of tests of main effects and interactions are presented in most textbook in a uniform manner, what follows after the significant interaction effects is considered to be a matter of confusion by several researchers (e.g., Jaccard, 1998, Kirk, 1995). Some aspects of the confusion are listed below:

1. Some researchers (e.g., Kirk, 1995) argue that SME analysis is deficient for revealing the nature of an interaction effect. Others (e.g., Jaccard, 1998, Keppel, 1991) maintain that SME analysis has its place since it is consistent with the theoretical framework researchers address and it is an easier way to interpret the interaction. An alternative approach to SME analysis is called “analysis of interaction comparisons” (Keppel, 1991) or “treatment-contrast interactions” (Kirk, 1995). This alternative approach decomposes the factorial design into smaller factorials and series of contrasts of interest are conducted. Detailed description of this approach can be found in Kirk (1995) and Keppel (1991).

2. Even within the SME analysis, there are some inconsistencies and ambiguities. First is the control of Type I error rates that involve multiple tests in factorial designs. Whether it is desirable or not, it is a common practice to ignore the effects of multiple tests involving the two main effects and interaction (in the two-way ANOVA). What about in the follow-up analyses? Most researchers recommend some type of control, but offer different approaches. To put in Lehman's (1995) term, “how to best proceed in following up a significant interaction is unclear. The active research literature seems to suggest ignoring the problem of the compounding Type I error” (pp. 450-451). Lehman recommends the Bonferroni adjustment in which the alpha is divided by the number of multiple tests. Keppel (1995) provides two approaches, doing nothing or using the Bonferroni adjustment and then suggests that further discussion is needed in this area. Kirk (1995) describes two procedures for controlling the Type I error rates in the SME analysis, Dunn’s procedure and the simultaneous test procedures (See pp. 381-382). Jaccard (1998) seems to prefer the modified Bonferroni method based on Holm (1979). Jaccard mentions that the modified Bonferroni method has been shown to be superior to the traditional Bonferroni method in terms of maintaining the experimentwise error rate without sacrificing statistical power in factorial ANOVA (Kromery & Dickinson, 1995).

3. The second problem associated with the SME analysis is what to do after the significant SME when there are more than two levels in the independent variable. Keppel (1991) calls such analysis “simple comparisons”. Some researchers recommend a contrast analysis (e.g., Keppel, 1995), while others recommend multiple comparison procedures such as Tukey or Scheffe methods (e.g., Lehman, 1995).

4. The third problem associated with the SME analysis is the choice of error term. Although
most authors of textbooks who mention the error term in the SME analysis point out that the error term from the original factorial ANOVA should be used, the problem is that sometimes it is not mentioned. If not mentioned, students may run separate one-way ANOVAs in which the error term based on the subgroup may be used since it is often the default in the statistical software.

5. We could not find a sound theoretical basis for the APC approach, although this approach seems to be often reported in journal articles. It is a special case of testing contrasts. Instead of testing some planned contrasts, this approach tests all possible contrasts. Therefore it is exploratory in nature. The statistical power of this approach is unclear.

Although a considerable amount of research is needed to compare and contrast different approaches and evaluate each method described above, it is not our intention to do so in this paper. Our intention is to acknowledge the existing confusing status and how it leads to difficulty in teaching coupled with the type of students we teach.

2.2 Second Source: Students

We must now clarify the type of students we teach. They are graduate students in the College of Education in an urban university. Most students have a full time job and most courses are offered in the evening. Factorial ANOVA is taught in the second course in statistics and it is a required research core course for most of the students. Some students will be consumers of quantitative research and some will be researchers who may use Factorial ANOVA in their dissertations and beyond. Due to the mix of students, the content of the course is more conceptual than technical, and practical aspects are emphasized more than the theoretical ones. Factorial ANOVA problems are analyzed by using SPSS. The following is the list of problems associated with the difficulty in teaching the follow-up tests in factorial ANOVA.

1. Students often have difficulty in conducting SME analysis because they are not sure how to split the data and what is going to be the independent variable in the SME analysis. This may stem from the fact that they are still confused about the independent variable versus the dependent variable in addition to the confusion about the number of factors versus the number of levels within the factor. This problem can be easily alleviated by making sure that students understand these concepts when one-way ANOVA is introduced.

2. Faced with SME analysis, students are often not sure if they should look at the effect of Factor A at each level of Factor B, or look at the effect of Factor B at each level of Factor A. Researchers usually recommend conducting one of them, not both. For example, Keppel (1991) mentions “Usually, we will choose to analyze the set of simple effects that is the most “natural” or potentially revealing- manipulation that will be the easiest to explain“ (p. 238). Jaccard (1998) proposes two terms, a focal independent variable and a moderator variable. A focal independent variable is the variable that a researcher is most interested in comparing and is often manipulated (such as treatment A vs. Treatment B). A moderator variable is a variable by which a researcher believes that the effect of the focal variable on the dependent variable is
“moderated” (such as gender). Once the distinction is clear, it is easy to explain to students that SME analysis is conducted at each level of a moderator variable keeping the focal variable as an independent variable.

3. Students rely on SPSS, sometimes overly. Running SPSS has become relatively easy since the introduction of the Windows version. Testing the main effects and the interaction is just a matter of clicks of buttons. However, SME is not part of a button. At this point, students need a step by step guide as to how to pursue the SEM analysis. In addition, SME analysis includes some hand calculations since the original error term from the factorial ANOVA should be used in calculating the F statistic. The same is true of a post hoc test after a significant SME when there are more than two levels in the focal variable.

4. We have found that insufficient information is sometimes more dangerous than no information at all. As we will see in the following section, some textbooks do not explain SME analysis sufficiently. A brief introduction of SME may lead to erroneous use of SPSS.

In summary, we identified existing problems associated with teaching the follow-up tests in factorial ANOVA. For those who are interested in the technical aspects of this problem, we found Jaccard’s (1998) book very helpful. Keppel (1991) describes the SME analysis in the most understandable way.

3. Survey of Statistics Textbooks

We attempted to investigate how introductory and intermediate statistics textbooks in education and psychology handle the problems discussed in the previous section. It was our speculation that most books handle the problems very cursory, since that was the case for the textbooks we have used in our course.

The selection of textbooks was limited to availability of the books and we do not claim that the books included here represent a random sample of statistics books. We limited the books included based on the following criteria:

1. Published after 1990.
2. Includes Factorial ANOVA.
3. Designed for social sciences (e.g., education and psychology).

Based on the problems identified in the previous section, the evaluation consisted of the following items:

1. Type of textbooks coded by the following criteria:
   A. too basic for our students (i.e., the second course in statistics)
   B. potential textbooks for our students but covers both first and second courses of statistics.
   C. potential textbooks for our students but targeted for the second course in statistics
D. too technical for our students
E. textbooks on how to use SPSS

2. Whether the follow-up tests are mentioned (Graph, SME, APC, or others)

If SME is included,
3. Whether the appropriate error term (MSE) is mentioned
4. Whether the control of Type I error ($\alpha$) is mentioned
5. Whether a post hoc test is mentioned in dealing with more than two levels in the focal variable.

Table 1 presents the summary of evaluation. The books are listed by the alphabetical order of author(s)'s name within each type.

An evaluation of selected textbooks confirmed our speculation that introductory and intermediate textbooks tended to include only graphing as a strategy for the follow-up tests. Even when SME analysis is included in their discussion, some of the aspects of SME such as the choice of error term, control of Type I error rates, and post hoc tests are missing, making the SME analysis incomplete. We found two of the textbooks (Keppel, 1991; Kirk, 1995) complete in their discussion of the follow-up tests. Unfortunately, those two textbooks are written for advanced researchers, and we find them not adoptable for our course.

Considering the lack of complete explanation of the follow-up tests in factorial ANOVA in most textbooks, the attached handout was developed to help students run the follow-up tests in factorial ANOVA using SPSS. Our intended audience includes graduate students who are in education or psychology and have a basic understanding of factorial ANOVA and SPSS. Technical and computational aspects are kept at a minimum. We chose approaches that are considered to be typical, since even statisticians do not agree on one approach. We presented the approach as one of many approaches, recommending further readings for other approaches. Within the selected approach we tried to be complete so that students can actually conduct his or her research until the very end of the follow-up tests.

The handout is available on the web at:
http://www.gsu.edu/~epstco/
<table>
<thead>
<tr>
<th>Textbook</th>
<th>Mention of Follow-Up Tests</th>
<th>If SME Is Mentioned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Type&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Graph</td>
</tr>
<tr>
<td>Hopkins et al. (1995)</td>
<td>A</td>
<td>✓</td>
</tr>
<tr>
<td>Hurlburt (1994)</td>
<td>A</td>
<td>✓</td>
</tr>
<tr>
<td>Kurtz (1999)</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>Lockhart (1998)</td>
<td>A</td>
<td>✓</td>
</tr>
<tr>
<td>Minium et al. (1999)</td>
<td>A</td>
<td>✓</td>
</tr>
<tr>
<td>Spatz (2001)</td>
<td>A</td>
<td>✓</td>
</tr>
<tr>
<td>Abrami, et al. (2001)</td>
<td>B</td>
<td>✓</td>
</tr>
<tr>
<td>Bakeman (1992)</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>Glass &amp; Hopkins (1996)</td>
<td>B</td>
<td>✓</td>
</tr>
<tr>
<td>Gravetter &amp; Wallnau (2000)</td>
<td>B</td>
<td>✓</td>
</tr>
<tr>
<td>Howel (1999)</td>
<td>B</td>
<td>✓</td>
</tr>
<tr>
<td>Huck (2000)</td>
<td>B</td>
<td>✓</td>
</tr>
<tr>
<td>Lehman (1995)</td>
<td>B</td>
<td>✓</td>
</tr>
<tr>
<td>Sprintal (2003)</td>
<td>B</td>
<td>✓</td>
</tr>
<tr>
<td>Boniface (1995)</td>
<td>C</td>
<td>✓</td>
</tr>
<tr>
<td>Lomax (1992)</td>
<td>C</td>
<td>✓</td>
</tr>
<tr>
<td>Stevens (1999)</td>
<td>C</td>
<td>✓</td>
</tr>
<tr>
<td>Keppel (1991)</td>
<td>D</td>
<td>✓</td>
</tr>
<tr>
<td>Kirk (1995)</td>
<td>D</td>
<td>✓</td>
</tr>
<tr>
<td>Kinnear &amp; Gray (1995)</td>
<td>E</td>
<td>✓</td>
</tr>
<tr>
<td>George &amp; Malley (2001)</td>
<td>E</td>
<td>✓</td>
</tr>
</tbody>
</table>

<sup>a</sup> Type A - Basic, Type B - Covers 1st and 2nd courses, Type C - Covers 2nd course only, Type D - Technical, Type E - SPSS books  
<sup>b</sup> Simple main-effects  
<sup>c</sup> All pairwise comparison
References


