
METHODOLOGICAL ARTICLE

Analysis of Variance Frameworks in Clinical Child and Adolescent Psychology: Issues and Recommendations

James Jaccard

Department of Psychology, State University of New York, Albany

Vincent Guilamo-Ramos

School of Social Work and Department of Population and Family Health, Columbia University

Reviewed existing practices of factorial analysis of variance (ANOVA), a major analytic tool used in clinical child and adolescent psychology, in the Journal of Clinical Child Psychology (JCCP) and noted several suboptimal strategies. Issues surrounding the analysis of multiple outcome variables, omnibus F tests, and single degree of freedom contrasts, simple main effects analysis, and single degree of freedom interaction contrasts were considered and recommendations were made about analytic strategies. Among the practices questioned were the use of multivariate analyses of variance (MANOVAs) as a means of controlling Type I errors across multiple outcome variables and the use of simple main effects analysis to elucidate the nature of interaction effects.

A wide variety of research methods are used in studies of children and adolescents in clinical psychology. Not surprisingly, diverse statistical methods are applied to the data yielded by these investigations. Statistical analysis has evolved considerably over the years, with social scientists having a broader array of statistical techniques than ever before to draw on. In addition, questions about the utility and viability of traditional null hypothesis testing (e.g., Cohen, 1994) have forced many investigators to re-think their approaches to data analysis. Despite these new perspectives and approaches, many social scientists continue to analyze their data using traditional approaches that have served the profession over the past decades. We conducted an analysis of research designs and statistical methods used in the *Journal of Clinical Child Psychology*¹ (JCCP) during the past 3 years. In many cases, we observed statistical practices that were either contrary to what should be done based on current statistical knowledge or where analyses could be augmented effectively with analytic methods that might prove enlightening. This state of affairs is not uncommon to other substantive domains beyond clinical child

and adolescent psychology. The purpose of this article is to alert researchers to selected analytic issues that should be considered when using analysis of variance (ANOVA) for designs that typically occur in clinical child and adolescent psychology. A wide array of statistical methods other than ANOVA were reported in JCCP, including multiple regression, structural equation modeling, growth curve analysis, factor analysis, and logistic regression, to name a few. Consideration of statistical practices for these latter methods is beyond the scope of this article. We amplify on practices that were used by at least one and usually more than one investigator. We do not provide formal citations to the studies because we do not see doing so as being constructive.

Traditional Approach

To make matters concrete, we refer to a hypothetical study that used a 3×2 factorial design where the first factor was the type of treatment to which children were randomly assigned (a cognitive therapy vs. a behavioral therapy vs. a control group) and the second factor was a theoretically mandated age grouping of the children participating in the study (younger vs. older). Of interest is whether there are differences among the three treatment groups on the primary outcome vari-

Requests for reprints should be sent to James Jaccard, Department of Psychology, University at Albany, State University of New York, Albany, NY 12222. E-Mail: jjj20@albany.edu

¹As of this issue, *Journal of Clinical Child Psychology* is now *Journal of Clinical Child and Adolescent Psychology*.

able (depression), as well as four secondary outcome variables (anxiety, fear reactions, quality of peer relations, and quality of family relations).

In the traditional approach, the investigator first conducts a 3×2 factorial multivariate analysis of variance (MANOVA) on the five outcome variables. The MANOVA is conducted so as to control for chance effects (i.e., inflated Type I error rates) across the five outcome variables. If the multivariate F test for a given factor is statistically significant, then univariate F tests for that factor are examined for each outcome variable individually. At the univariate level, the researcher may examine the F tests for the main effects or the interaction effect. If the univariate F test for a main effect is statistically nonsignificant, then no further analyses for that effect are pursued. If the univariate F test for the main effect of the treatment factor is statistically significant, then follow-up analyses are pursued to determine which of the three group means are statistically significantly different from one another (because the treatment factor has more than one degree of freedom). This takes the form of conducting statistical comparisons between all possible pairs of means (i.e., comparing the cognitive treatment group with the control group, the behavioral treatment group with the control group, and the cognitive treatment group with the behavioral treatment group). Because multiple follow-up tests are performed, a procedure is invoked to control the experimentwise error rate across the pairwise contrasts. The choice of method to accomplish this varies, with the most popular ones being the Tukey honestly significant difference (HSD) method or a Bonferroni correction. Based on the results of these pairwise contrasts, conclusions are made about the population mean differences among the three groups. Because age has only two levels, no follow-up tests are required if a statistically significant main effect for age is observed.

If a statistically significant F test for the interaction is observed, then follow-up analyses are performed to elucidate the nature of the interaction. The most common strategy is to apply simple main effects analysis. This involves conducting a one-way ANOVA on the outcome measure as a function of the three treatment groups focusing first on only the younger children (but using the error term from the overall 3×2 ANOVA in the denominator of the F test). If this F test for the simple main effect is statistically significant, then follow-up analyses involving pairwise mean comparisons are pursued using a Tukey test or Bonferroni-based test to control for chance effects across the pairwise comparisons. These simple main effect analyses are then repeated focusing only on the older children.

This analytic approach, though common, has pitfalls. In this article, we examine issues of (a) the analysis of multiple outcome variables, (b) the conduct of "follow-up" tests on factors that have three or more levels, and (c) the exploration of interaction effects. In

a companion article to this one (Jaccard & Guilamo-Ramos, in press), we discuss more advanced (but equally important) issues in ANOVA and consider designs with repeated measures. Our general strategy is to divide each section into three subsections. The first subsection explicates the essence of the issue and the inadequacy of common practice. The second subsection provides a commentary in which the underlying statistical issues are explored on a more technical (but still readable) level. This section can be skimmed by readers who are less interested in statistical details. The final subsection discusses action steps that the analyst should consider implementing in practice. We provide such recommendations with some trepidation, knowing that there may be exceptions in which an alternative analytic strategy would better answer the researchers questions. However, the strategies we suggest should find widespread applicability in problems traditionally studied by child and adolescent psychologists.

Multiple Outcome Measures

Nature of the Problem

The majority of investigations that we examined in *JCCP* that used ANOVA-based strategies had multiple outcome variables. It is well known that when multiple outcomes are analyzed in separate factorial ANOVAs, the chances of falsely concluding that an effect exists inflates across the analyses. Specifically, if an alpha level of 0.05 for a factor is used for each outcome variable separately, then although the Type I error rate for a given analysis is indeed 0.05, the probability that at least one of the analyses across the five outcomes will yield a Type I error exceeds 0.05, sometimes substantially so. The problem for the investigator is to control the error rate across the multiple outcomes and to do so in a way that does not increase by too much the risk of missing an effect that truly exists. The control for chance effects across outcomes is what typically motivates investigators to adopt the MANOVA strategy. However, the analytic strategy is inadequate for doing so.

Commentary

Why does the probability of chance effects increase across multiple outcomes? Statisticians distinguish between two types of error rates. The first is a *per comparison* error rate, which refers to the probability of falsely rejecting the null hypothesis for a single significance test. For example, if we compare the cognitive treatment group with the control group on depression and falsely conclude that a population difference in means exists when, in fact, it does not, then a Type I error has occurred for this particular comparison. The proba-

bility of a Type I error for a given comparison is the alpha level, traditionally 0.05. The second error rate is called the *experimentwise error rate* (although it goes by other names as well) and refers to the error rate across multiple contrasts, multiple comparisons, or multiple significance tests. For example, if the cognitive treatment group is compared with the control group on each of the five outcome measures, then five comparisons have been performed. The rate at which at least one chance effect occurs across the five comparisons is the experimentwise error rate. Even if the per comparison error rate is 0.05, the experimentwise error rate across the multiple comparisons will be larger than 0.05.

Consider a simple coin-flipping analogy. If we flip a coin, there are two possible outcomes that can occur, one of which is a “head.” The likelihood of observing a head on a given coin toss is thus $1/2 = 0.50$. If we flip a coin twice, there are four possible outcomes that can occur: (a) a head on the first flip followed by a head on the second flip, (b) a head on the first flip followed by a tail on the second flip, (c) a tail on the first flip followed by a head on the second flip, (d) a tail on the first flip followed by a tail on the second flip. Note that a head occurs on three of the four flips, so the probability of a head occurring on at least one of the flips is $3/4 = 0.75$. Even though the probability of a head is 0.50 on a given flip, the probability of observing at least one head across two flips is 0.75. Psychologists traditionally desire to invoke analytic procedures that will maintain a 0.05 error rate across multiple comparisons, hence the use of the MANOVA strategy.

Why doesn't the MANOVA strategy always work? Statisticians distinguish between the concepts of a complete null hypothesis and a partial null hypothesis. The complete null hypothesis is true when there are no group differences in population means on any of the outcome variables. A partial null hypothesis is true when there are group differences in population means for some of the outcome variables but not for others. The MANOVA strategy adequately controls the experimentwise error rate across the multiple outcome variables when the complete null hypothesis is true. However, if the partial null hypothesis is true, MANOVA tends not to do so.

Consider the case where a partial null hypothesis is operating for the main effect of the treatment factor in our example. Suppose that there are large population group differences on depression but no such population group differences on the other four outcome variables. The large group differences on depression probably will result in the rejection of the null hypothesis for the multivariate F test of group differences on the outcome means, considered simultaneously. Once this rejection has taken place, the investigator proceeds to conduct five univariate F tests on each of the outcome variables. At this point, the five F tests are not subject

to any form of experimentwise control, and the Type I error rate across the four outcome measures that have no population differences inflates beyond the traditional 0.05 level, much like in the coin-flipping analogy. Had the depression measure (where the large population differences exist) not been included in the analysis, then the complete null hypothesis would be true, and MANOVA would indeed control for an inflated error rate across the outcome variables (assuming all other assumptions of the technique were met). However, when the partial null hypothesis is true by virtue of including the depression measure, this is not necessarily the case.² In general, if the concern is with controlling the experimentwise error rate across multiple outcome measures, then an omnibus MANOVA F test used as a “screen” for traditional univariate analyses is unsatisfactory for doing so because, in practice, we are probably operating in scenarios where the partial null hypothesis is true.

Can we apply MANOVA first and then use a formal method for controlling experimentwise error rates rather than uncorrected univariate ANOVAs? Some investigators perform omnibus MANOVA F tests and then, given a statistically significant multivariate F test, conduct follow-up analyses on each outcome invoking a formal experimentwise error control procedure across the outcome variables. Instead of relying on univariate F tests for the follow-up analyses, the investigator applies simultaneous F tests following the strategy outlined in Morrison (1959) or a Bonferroni correction across the univariate F tests. These practices are problematic because most of the control procedures were not devised with the idea that there would be a two-step process involving an initial screening test using the overall multivariate F test. Rather, the methods were developed without regard to the results of a multivariate F test and should be applied independent of the results of such tests.

What is a better strategy for controlling the experimentwise error rate? If one desires to control for inflated Type I error rates across multiple outcome variables, then one of the better strategies for doing so is to forgo MANOVA and conduct separate univariate F tests on each outcome variable. A Bonferroni-based correction can then be applied across the outcome variables. Most investigators who adopt this approach apply the traditional Bonferroni correction by dividing the per comparison alpha level (usually 0.05) by the

²One cannot know if the partial or complete null hypothesis is actually true based on the results of the univariate F tests. Scenarios can exist where the partial null hypothesis is true but all of the univariate F tests are statistically significant or where the partial null hypothesis is true and none of the univariate F tests are statistically significant. The same is true for the complete null hypothesis.

Table 1. Statistical Power for Selected Alpha Levels as a Function of Per Group Sample Size and Population Effect Size

	Number of Outcome Variables				
	1 ($\alpha = .050$)	2 ($\alpha = .025$)	3 ($\alpha = .0167$)	4 ($\alpha = .0125$)	5 ($\alpha = .010$)
Small effect size					
20 per group	.09	.05	.03	.03	.02
30 per group	.12	.07	.05	.04	.03
50 per group	.17	.10	.07	.06	.05
70 per group	.22	.14	.11	.09	.07
100 per group	.23	.20	.16	.13	.12
Medium effect size					
20 per group	.34	.23	.19	.16	.14
30 per group	.48	.36	.30	.26	.24
50 per group	.70	.58	.52	.48	.45
70 per group	.84	.75	.70	.66	.63
100 per group	.94	.89	.86	.84	.82

number of outcome variables (in this case 5) and then use this as the critical alpha level for each univariate analysis. In our example, because there are five outcome variables, the critical alpha level would be $0.05/5 = 0.01$. Only if a univariate analysis yielded a p value less than .01 would it be declared statistically significant. This particular Bonferroni strategy is overly conservative and often has low statistical power. Alternative Bonferroni-based methods are available that have higher levels of statistical power. Appendix A describes two such procedures that are generally preferable to the traditional Bonferroni method.

Should controls for experimentwise error rates routinely be invoked with multiple outcome variables? Many investigators feel that controls for inflated experimentwise error rates should be invoked whenever multiple outcomes are present. However, the issue is more complex than this. Using such controls reduces statistical power for a given comparison, with the result possibly being an unacceptably high rate of Type II errors. Using the Bonferroni procedure in our example, if the critical alpha level for a contrast to be declared statistically significant is set at 0.01 rather than 0.05, then obviously we will be less likely to declare any given result statistically significant and we run a higher risk of missing an effect that is real. If the error of failing to detect an effective treatment is a serious one, then one must carefully balance this possibility against controlling for chance effects across multiple comparisons. In paradigms where sample sizes tend to be small, the issue is particularly germane because statistical power is low to begin with. In our review of studies that used ANOVAs published in *JCCP*, sample sizes were often in the range of 30 to 40 participants per cell. Table 1 presents the probability of detecting a mean difference between two groups (treatment group vs. control group) if one were to apply a traditional Bonferroni correction for up to five outcome variables using small and medium population ef-

fect sizes as defined by Cohen (1988). It can be seen that statistical power tends to be low even for a single outcome variable for sample sizes typical of clinical child research and that applying an experimentwise error rate correction across multiple outcomes only exacerbates an already bleak situation in terms of Type II errors. In such situations, one might decide not to invoke the experimentwise controls because the effect on statistical power is too severe and conceptually costly.

How can we balance the need for high statistical power and the need to control the experimentwise error rate?

The most common way of balancing Type I and Type II errors is to define different “families” of outcome variables where the experimentwise error rate is controlled within a family but not across families. A family of outcomes is a subgroup of outcomes that have been grouped together based on theoretical or practical criteria. In this strategy, the researcher makes a point of controlling for experimentwise error *within* a family of outcomes but sacrifices such controls across families in the interest of statistical power. As an example, in the treatment study, a researcher could invoke experimentwise error rate controls across all five outcome variables. Alternatively, he or she might group the outcome variables into two families, a family of primary outcomes (depression) and a family of secondary outcomes (anxiety, fear reactions, quality of peer relations, and quality of family relations). For the first family, there is only a single outcome variable, so no experimentwise error control procedure is necessary. For the second family, there are four outcome variables, and experimentwise error controls are invoked across these four outcomes. This approach will have more statistical power than the first one that controlled for experimentwise error across all five outcomes, but it comes at some cost over less control of Type I errors.

The strategy of defining families is commonplace in data analysis in psychology. Doing so yields three Type I error rates that are of potential interest: (a) the

per comparison Type I error rate, which is the probability of making a Type I error for a given outcome; (b) the within-family Type I error rate, which is the probability of making at least one Type I error across the multiple outcomes within a family; and (c) the across-family Type I error rate, which is the probability that at least one of the multiple families will have a Type I error. Statisticians use different labels for these concepts, which can be confusing. However, the essence of the three types of error rates is straightforward.

How does one decide what variables to group together into a family? A difficult issue is one of specifying the criteria for grouping outcomes into families for purposes of invoking experimentwise controls. For example, one might decide to group into a family all those measures that reflect the same construct, such as three different measures of depression. This strategy makes it more difficult to reject the null hypothesis for any one of the depression measures because the per comparison alpha level for a single outcome will be lower when the experimentwise error controls are invoked. This, in turn, lowers the statistical power for detecting an effect on the concept of depression. So this approach is not necessarily desirable.³

The grouping of families often will be dictated by theoretical concerns, which we cannot anticipate here because such concerns will vary with each substantive domain. It is our experience that most child clinical studies have one or two primary outcome variables and multiple secondary outcome variables. One analytic strategy might be to define two families of contrasts, one based only on the primary outcome variables and then a second family based only on the secondary outcome variables. In our 3×2 intervention example, the first family consists of a single variable, depression, and the second family consists of the four secondary variables: anxiety, fear reactions, peer relations, and family relations. Because the first family has only one variable, the traditional 0.05 alpha level is applied. Because the second family has four outcome variables, the Bonferroni strategy described in Appendix A is applied to them (setting k equal to 4). Some methodologists argue that exploratory analyses should not adopt more strict controls because the spirit of such analyses is to isolate theoretical leads that are worthy of more rigorous study in future research. This philosophy would dictate against invoking the experimentwise controls in the family of secondary outcome variables, as long as the statistically significant results are viewed as tentative and suggestive of further study. Other methodologists argue that exploratory analyses should adopt more strict controls because they have less theo-

retical justification and one should discourage “fishing for effects.” Both positions have merit. There is no one best, unambiguous way to define families. The topic has been debated by statisticians for many years. An excellent source on the topic is Kirk (1995).

Action Steps

Use of MANOVA as a screen to follow-up univariate F tests usually is a poor strategy for controlling experimentwise error rates across multiple outcome variables. Bonferroni-based procedures in conjunction with univariate F tests are preferable.

Invoking controls for chance effects across multiple outcome variables should not be done routinely. Such controls lower statistical power, making it more likely that an investigator will miss an effect that may be important. If sample sizes are large and statistical power is high, then the controls can be readily invoked. However, there may be situations where the researcher simply does not want to sacrifice statistical power for the sake of controlling chance effects across outcome variables.

Researchers need to carefully think about the practical and theoretical consequences of concluding that an effect does not exist when, in fact, it does, versus the practical and theoretical consequences of concluding that an effect does exist when, in fact, it does not. Is it worse to overlook an intervention protocol that effectively treats depression or to conclude that an intervention protocol effectively treats depression when it does not? If both types of errors are deemed equally important, then why adopt traditional analytic strategies (e.g., using an alpha level of 0.05; invoking experimentwise controls across multiple contrasts) that, in practice, give precedence to the avoidance of Type I errors.

An obvious solution to the trade off of Type I versus Type II errors is to use large sample sizes in one's research, so that the issue of reduced statistical power when experimentwise controls are invoked is moot. In such cases, statistical power remains high even when the controls are applied. For small sample size research, the dilemma must be confronted directly. The solution might require raising one's per comparison alpha level above 0.05 rather than lowering it by invoking controls for experimentwise error rates. Readers should not take our statements to mean that one should casually raise per comparison alpha levels beyond the traditional 0.05 level. Rather, we are encouraging researchers to think about the consequences of both Type I and Type II errors and to adopt appropriate analytic strategies in light of such considerations. Adherence to the traditional alpha level of 0.05 in research situations where power is less than 0.95 is implicitly taking a stand on the issue: You are declaring that a conclusion

³Analytic strategies for multiple indicators of the same construct are discussed in Hancock, Lawrence, and Nevitt (2000).

that a treatment is effective when it is not is worse than missing an effective treatment.

One useful analytic approach for multiple outcome scenarios is to conduct analyses both with and without experimentwise controls within a family. If the results of the analyses for both forms of analysis are comparable in terms of declarations of statistical significance, then the issue of decreased power as a function of invoking experimentwise error rate controls is moot. If the results conflict, so that statistical significance is declared without the controls but not when the controls are invoked, then one must approach conclusions more cautiously and with theoretical tentativeness.

Omnibus F Tests and Single Degree of Freedom Contrasts

Nature of the Problem

When analyzing a single outcome variable, a common practice for cases where factors have more than two levels is to use a two-step testing strategy: (a) examine the overall F test for the factor to determine if it is statistically significant, and (b) if the F test is statistically significant, conduct additional mean comparisons to elucidate those differences that contribute to the rejection of the overall null hypothesis. The latter comparisons usually take the form of pairwise comparisons among subgroups using the Tukey HSD procedure or a Bonferroni procedure. For example, the mean depression score for the cognitive intervention group is formally compared with the mean depression score for the control group as well as the mean depression score for the behavioral intervention group. In addition, the mean depression score for the behavioral intervention group is compared with mean depression score for the control group.

The follow-up tests after a statistically significant overall F test are often referred to as single degree of freedom contrasts. The pairwise comparison of means using the Tukey HSD test is an example of a set of single degree of freedom contrasts with an experimentwise control applied to them. In our example, the Tukey HSD procedure has three single degree of freedom contrasts focused on group means, the cognitive treatment group versus the control group, the behavioral treatment group versus the control group, and the cognitive treatment group versus the behavioral treatment group. Single degree of freedom contrasts are at the heart of most theoretical questions addressed in ANOVA designs. The fact that more than one such contrast is performed in an experiment raises the possibility of inflated error rates (chance effects) across the multiple contrasts. A fundamental issue is how best to control the error rate. In practice, using the initial F test as a screen before conducting a Tukey test

or before applying a Bonferroni correction is not appropriate. In addition, several methods that have been proposed for controlling experimentwise error rates (e.g., the Newman–Keuls procedure, the Duncan procedure) sometimes do so inadequately.

Commentary

What is a single degree of freedom contrast?

When a factor has more than two levels, the F test associated with that factor has more than one degree of freedom in the numerator. In our example, the treatment factor has three groups and, consequently, the F ratio used to test the overall null hypothesis has two degrees of freedom in the numerator. A single degree of freedom test or “contrast” is one that has only a single degree of freedom in the numerator of the F statistic. The age factor in our example has only two levels, with a single degree of freedom in the numerator of the F ratio associated with it. This F test represents a single degree of freedom contrast.

When a factor has more than a single degree of freedom, the overall F test for the factor (also called an omnibus F test) only provides information about the complete null hypothesis for all the groups comprising the factor. For example, in our intervention example, the null hypothesis for the treatment factor is that all of the intervention groups (the cognitive intervention group, the behavioral intervention group, and the control group) share a common population mean. If the omnibus F is statistically significant, we reject this null hypothesis as being untenable. But rarely would a researcher stop there. We want to probe further and determine exactly which group means differ significantly from one another. This is where single degree of freedom contrasts come into play. When we compare the means of two groups in a more focused test (by using, for example, a Tukey HSD procedure), we are conducting a single degree of freedom contrast. Indeed, it is usually the case that answers to theoretical questions only emerge in the context of the single degree of freedom contrasts. The omnibus test is simply “too omnibus” to answer our more focused questions. As will be seen shortly, single degree of freedom contrasts can take forms other than that of a comparison of two means.

Why does one perform an omnibus F test? The most common reason people give for conducting the omnibus F test is that it helps to control the experimentwise error rate for the multiple single degree of contrasts that follow. In essence, a two-step procedure is adopted in which a statistically significant omnibus test is followed by pairwise comparisons of means using the Tukey HSD test or the Bonferroni method. A problem with this strategy is that the Tukey method and the Bonferroni methods were designed to control experimentwise error rates

across a set of contrasts without the initial screening based on the omnibus F test. If researchers have a set of single degree of freedom contrasts that are of conceptual interest and desire to control the experimentwise error rate across the contrasts, then they should apply the Tukey HSD test or a Bonferroni-based method without regard to the omnibus test. Indeed, the omnibus test usually takes on secondary interest to the more focused single degree of freedom tests and may even be of no conceptual interest whatsoever.

There exist procedures for testing single degree of freedom contrasts other than the Tukey HSD test or the Bonferroni-based methods that require that an omnibus F test be used in a screening capacity to control for experimentwise error rates. In such cases, one obviously has been interested in the omnibus F test because it is an integral part of the analytic method. However, most of these approaches have fallen into disrepute because they only control the experimentwise error rate when the complete null hypothesis is true (i.e., when all of the different subgroups of a factor have equal population means). When a partial null hypothesis is true (i.e., when some of the subgroups have equal population means but others do not), the techniques often fail.

Researchers also may have interest in omnibus effects when a conceptual question specifically focuses on the behavior of the overall factor. For example, some investigators desire to document the percent of variance accounted for by a given factor (e.g., how much of the total variation does the age factor account for), which generally leads to theoretical statements based on the behavior of the overall factor.⁴ Despite this, most applications in clinical child and adolescent psychology ultimately focus on single degree of freedom contrasts and, hence, the omnibus test is of lesser interest.

What method should be used for controlling experimentwise error rates across a set of single degree of freedom contrasts? The choice of a method for controlling experimentwise error rates across a set of single degree of freedom contrasts in factorial designs is complex, and we dare not attack this issue in all its glory here. Much has been written about the topic, and interested readers are referred to Toothaker (1993), Kirk (1995), Westfall, Tobias, Rom, Wolfinger, and Hochberg (1999), and Wilcox (1996). More than 30 such procedures have been proposed in the statistical literature. Technically, most multiple comparison approaches were derived for the purpose of comparing

means in single-factor designs. Extensions to factorial ANOVA is not always straightforward. The Bonferroni-based method in Appendix A is flexible in that it applies to cases of equal or unequal group sample sizes, cases of one-way designs or factorial designs, cases where statistical corrections are introduced for assumption violations, cases involving between-participant or within-participant factors, and cases where either all possible pairwise contrasts are of interest or only a subset of contrasts are of interest. The major disadvantages of the approach are that (a) there may be more specialized alternatives that have higher statistical power given the specific features of the research design and (b) the calculation of confidence intervals is undermined (see Jaccard & Guilamo-Ramos, in press).

In our review of the *JCCAP*, we observed the use of experimentwise error rate control procedures that have been rejected by statisticians as either providing inadequate control of experimentwise error rates (e.g., Duncan tests, Newman-Keuls tests) or as being overly conservative for the type of analysis being pursued (e.g., the traditional Bonferroni method, the Scheffé method). When group sample sizes are equal and the focus is on all possible pairwise comparison of means, the Tukey HSD test has served the profession well. With unequal sample sizes and, again, when the focus is on all possible pairwise comparisons, the Tukey-Kramer method is often recommended and has many merits. These procedures (as well as more than 20 additional such methods) are discussed in Kirk (1995). In situations where low power is a concern, there are alternative methods to the two Tukey procedures that can yield some gain in statistical power under certain conditions. These methods are discussed in Kirk and Westfall et al. (1999). The previously mentioned recommendations are based on scenarios where the statistical assumptions underlying ANOVA are met. Robust methods are discussed in Westfall et al. and Wilcox (1996, 1997, 2001).

What are the families for single degree of freedom contrasts? Traditional practice for applying experimentwise error controls in factorial ANOVA is to let each factor define a family of contrasts. In our research example, the treatment main effect represents one family, and it has three pairwise mean contrasts within it (the comparison of the cognitive treatment with the control group, the comparison of the behavioral treatment with the control group, and the comparison of the cognitive treatment with the behavioral treatment). The age main effect represents a second family that has only one single degree of freedom contrast within it (younger vs. older). The Treatment \times Age interaction is a third family. We discuss the single degree of freedom contrasts for the interaction effect in a later section. Most researchers do not invoke controls for experimentwise errors across families (i.e., across

⁴The tests also may be of interest if one adopts a variance decomposition philosophy (see Jaccard, 1998, and Kirk, 1995). Also, if the F test for an overall factor is not even remotely close to statistical significance, then it can serve as an effort-saving device to calculating the single degree of freedom contrasts.

the main effects and the interaction effect), but do so within a family.

Are there times when one does not want to invoke experimentwise controls across multiple single degree of freedom contrasts? The answer to this question draws on the same issues discussed earlier with respect to multiple outcome variables. Although controlling experimentwise error rates is desirable, it can be costly in terms of statistical power. If sample sizes are large, then the controls can be invoked with minimal adverse effects. When sample sizes are small to moderate, one must balance the consequences of making Type II errors against the additional control of chance effects. Scenarios may arise where one does not want to sacrifice statistical power to the extent required by the experimentwise control procedures.

It is sometimes asserted that controls for experimentwise error rates should be invoked for exploratory contrasts but not for a priori or planned contrasts as dictated by theory. Consider the following example: Two researchers conduct the identical experiment that uses a one-way design with an independent variable that has three groups. The first investigator has a theory that leads her to posit three a priori contrasts that correspond to comparing Group 1 versus Group 2, Group 1 versus Group 3, and Group 2 versus Group 3 (i.e., she plans to conduct all possible pairwise contrasts). The second investigator does not have such a theory and, instead, adopts an exploratory mode in which he plans to examine all possible pairwise contrasts. According to some, the first investigator would not invoke statistical procedures to control for the experimentwise error rate because the contrasts are “planned,” whereas the second investigator should invoke such controls. Note that the only difference in the two scenarios is that the first investigator has uttered a set of words about a theory before conducting the experiment. However, “chance” does not have ears. It does not matter if the first investigator says “I have such and such a rationale for doing this particular contrast” any more than if the investigator says “I think I will eat meat for dinner tonight.” Chance will be operating in both situations, no matter what words come out of a person’s mouth. The central issue is not whether a set of contrasts are exploratory or guided by theory. Rather, the issue is whether one desires to control the experimentwise error rate.

Action Steps

Researchers should specify a priori the single degree of freedom contrasts that are of interest in a given study and then evaluate the statistical significance of those contrasts. For the main effects in the intervention study, there are three single degree of freedom contrasts associated with the treatment factor that are of interest (cognitive group vs. control group, behavioral

group vs. control group, and cognitive group vs. behavioral group) and one single degree of freedom contrast associated with the age factor (younger vs. older). Each factor defines a family of single degree of freedom contrasts. Given sufficient statistical power, the researcher should control the experimentwise error rates within a family. For equal sample sizes per group and when interest is in all possible pairwise comparisons, a useful procedure is the Tukey HSD test. If there are unequal sample sizes, the Tukey–Kramer method can be used. There is no need to use the omnibus F test as a screen in these approaches. Many of the early methods proposed by statisticians for controlling experimentwise error rates do not do so successfully. Others are overly conservative. Given that the underlying statistical assumptions of ANOVA are met, the aforementioned methods will serve the analyst well. Other approaches that can be considered are described in Kirk (1995), Westfall et al. (1999), and Wilcox (1996, 2001).

For studies with small to moderate sample sizes, it will probably prove useful to pursue the various single degree of freedom contrasts both with and without experimentwise controls to determine if conclusions are comparable across the two analytic strategies. If conclusions are the same with and without the controls, then one can move forward accordingly. If the conclusions change depending on whether the controls are invoked, then one should consider approaching his or her conclusions with theoretical tentativeness.

Single Degree of Freedom Interaction Contrasts

Nature of the Problem

The traditional approach to analyzing a statistically significant interaction effect in a factorial design is to use simple main effects analysis. Simple main effects address the question of whether an independent variable affects a dependent variable at each level of a moderator variable. For example, we might ask if the treatment subgroups differ in their population means when just the younger children in the intervention study are considered. We might also ask whether the treatment subgroups differ in their population means when we consider just the older children. These are meaningful questions and ones that an analyst typically wants to examine. However, they have little bearing on an interaction effect. To effectively explore interactions, one must pursue single degree of freedom interaction contrasts (Boik, 1979; Jaccard, 1998).

Commentary

What is an interaction effect? Interaction effects can be parameterized (or thought about) in differ-

ent ways. One of the most common conceptualizations invokes the notions of a dependent variable, an independent variable, and a moderator variable (Holmbeck, 1997; Jaccard, 1998). The independent variable is the presumed cause of the dependent variable. In the case of an interaction, the effect of the independent variable on the dependent variable changes depending on the value of a third variable, called the moderator variable. For example, the effect of the cognitive intervention relative to the control group might be stronger for younger children than it is for older children. This differential effect as a function of age is the essence of an interaction effect.

Suppose that for younger children, the mean depression score for the cognitive treatment group at the immediate posttest was 1.98 (on a 1 to 5 scale), whereas for the control group it was 3.80. The effect of the cognitive treatment can be quantified as $1.98 - 3.80 = -1.82$. That is, relative to the control group, the cognitive intervention lowered depression scores by 1.82 units for the younger children. What about for older children? Suppose that the mean depression score for older children in the cognitive treatment group was 1.96 and for the control group it was 2.10. For older children, the effect of the cognitive intervention is quantified as $1.96 - 2.10 = -.14$. That is, relative to the control group, the cognitive intervention lowered depression scores by .14 units.

Are the effects of the cognitive intervention relative to the control group the same for the two age groups? It seems not. For younger children, the intervention lowered depression scores by 1.82 units whereas for older children it did so by 0.14 units. We can quantify how disparate the two effects are by taking the difference between them. The difference between -1.82 and $-.14$ is -1.68 . The cognitive intervention (relative to the control group) lowered depression scores an average of 1.68 units more when the intervention was directed at younger children than when it was directed at older children. This single number, -1.68 , is called an *interaction parameter estimate* because it is an estimate of how much stronger the effect of an independent variable is in one group (younger children) as compared to another group (older children).

Because we are dealing with sample data, it is possible that the true interaction parameter in the population is zero (i.e., that the effects of the cognitive intervention relative to the control group are identical for younger and older children) and that we have observed an interaction parameter value of -1.68 just by chance. We will want a test of a null hypothesis of a zero interaction parameter to evaluate if a value of -1.68 is unlikely to occur by chance when the null hypothesis is true. The statistical test that accomplishes this has a single degree of freedom, hence it is a single degree of freedom interaction contrast.

Single degree of freedom interaction contrasts are fundamental to interaction analysis. Whenever a statis-

tically significant interaction is observed, it implies that the effect of the independent variable on the dependent variable changes depending on the value of the moderator variable. The interaction parameter quantifies in a precise fashion how the effect in one group varies from that in another group. The value of -1.68 indicates how much more effective the cognitive treatment was in younger children as compared with older children.

Why do simple main effects not address the issue of interaction?

As noted, an interaction contrast compares the effect of an independent variable on a dependent variable in one group with the comparable effect in another group. By contrast, simple main effects make no such comparisons. A simple main effect focuses on only one group defined by the moderator variable and asks if the independent variable had an effect for that particular group. For example, did the cognitive intervention lower depression scores relative to the control group for just the younger children? Our data suggested that this was the case because the mean for the cognitive intervention for the younger children was 1.98, whereas the mean for the control group for the younger children was 3.80, a difference of -1.82 . But note that a statistical test of the significance of this simple main effect fails to compare the effect with any other group (i.e., it does not compare the effect with that for older children). It does not address the issue of statistical interaction.

We can illustrate the point another way using a correlation example. Suppose that two variables, X and Y , are correlated 0.24 for boys and 0.22 for girls. Suppose that the correlation is statistically significant for boys ($p < .05$) but not for girls ($p > .05$). These significance tests within each group, are analogous to simple main effect tests. Can we conclude from these data that the correlation between X and Y is stronger for boys than it is for girls? Certainly not. Even though the correlation was statistically significant in one group but not the other group, we can only say that the correlations differ if we directly test the difference between the two correlations. This test of the difference between the two correlations (which in this case is not statistically significant) is analogous to a test of interaction.

Are simple main effect analyses ever of interest?

Absolutely. Most researchers will want to know whether the effect of an independent variable on a dependent variable is statistically significant at each level of the moderator variable. Simple main effects usually are of theoretical import. However, they do not bear on interaction effects, which address a different substantive question.

How would one approach interaction contrasts in our example study?

Because the omnibus interaction effect has two degrees of freedom, the analysis

is somewhat involved, yet conceptually simple. We begin by identifying the dependent variable, the independent variable, and the moderator variable. Depression is the dependent variable, the treatment condition is the independent variable, and age is the moderator variable. Table 2 presents hypothetical mean scores for each cell of the design.

Next we specify interaction-based questions that are of conceptual interest. There are three. First, is the difference in mean depression scores between the cognitive intervention group and the control group the same for younger and older children? Second, is the difference in mean depression scores between the behavioral intervention group and the control group the same for younger and older children? Third, is the difference in mean depression scores between the cognitive intervention group and the behavioral intervention group the same for younger and older children? Answering each of these questions will provide a sense of how the effects of the treatments differ as a function of the moderator variable.

As it turns out, each of these questions focuses on a different 2 × 2 subtable in the overall design, as shown in Table 2. Consider the first subtable. It addresses the first question. The effect of the cognitive intervention relative to the control group for younger children is indexed by the difference between the two means in the first column, 1.98 – 3.80 = –1.82. The effect of the cognitive therapy relative to the control group for the older children is indexed by the difference between the two means in the second column, 1.96 – 2.10 = –0.14. As discussed before, the difference between these two indexes is –1.68, and it indicates how different the effect of the cognitive intervention is (relative to the control condition) for younger children as opposed to older children.

The second 2 × 2 subtable addresses the second question. The effect of the behavioral intervention relative to the control group for younger children is indexed by the difference between the two means in the first column, 3.10 – 3.80 = –0.70. The effect of the behavioral therapy relative to the control group for the

older children is indexed by the difference between the two means in the second column, 2.02 – 2.10 = –0.08. The difference between these two indexes is –0.62, indicating that the behavioral intervention (relative to the control condition) lowered depression scores by 0.62 units more for younger children than for older children.

The third 2 × 2 subtable addresses the third question. The effect of the cognitive intervention relative to the behavioral intervention for younger children is indexed by the difference between the two means in the first column, 1.98 – 3.10 = –1.12. The effect of the cognitive therapy relative to the behavioral therapy for the older children is indexed by the difference between the two means in the second column, 1.96 – 2.02 = –0.06. The difference between these two indexes is –1.06, indicating that the cognitive intervention relative to the behavioral intervention lowered depression scores by 1.06 units more for younger children than for older children.

Having identified the interaction parameters of interest, the final step is to conduct significance tests of them. Methods for doing so are presented in Appendix B. This appendix also discusses how to conduct tests of simple main effects for the example.

In sum, the analysis of an interaction effect involves specifying single degree of freedom interaction contrasts that are of conceptual interest. This usually takes the form of specifying 2 × 2 subtables of the overall design and then calculating the value of the interaction parameter estimate for each subtable. A test of significance is then applied to the parameter estimate. Few studies reported in *JCCP* focused on single degree of freedom interaction parameters. Rather, interaction analysis was typically undertaken by simple visual inspection of the means or by misapplication of simple main effects analysis. Specifically, if a simple main effect was statistically significant for one group (e.g., boys) but not another group (e.g., girls), then the interaction effect was said to be evident. As noted earlier, this logic fails to formally compare effects in the two groups. Single degree of freedom interaction contrasts make such comparisons.

Table 2. Hypothetical Depression Means as a Function of Treatment Group and Age

	Younger	Older	Marginal Mean
Cognitive	1.98	1.96	1.97
Behavioral	3.10	2.02	2.56
Control	3.80	2.10	2.95
Marginal mean	2.96	2.03	

	Younger	Older
Cognitive	1.98	1.96
Control	3.80	2.10
	(1.98 – 3.80) – (1.96 – 2.10)	
	–1.68	

	Younger	Older
Behavioral	3.10	2.02
Control	3.80	2.10
	(3.10 – 3.80) – (2.02 – 2.10)	
	–0.62	

	Younger	Older
Cog	1.98	1.96
Behavioral	3.10	2.02
	(1.98 – 3.10) – (1.96 – 2.02)	
	–1.06	

Single degree of freedom interaction parameter estimates are not vague statistical abstractions. They map onto vitally important conceptual questions, and a fundamental understanding of their nature is essential for effective interaction analysis. For a more detailed discussion of them and several numerical examples, see Jaccard (1998) and Boik (1979).

How does one identify the moderator variable in a factorial design? The decision about which factor should assume the role of “moderator” status in a factorial design is based on conceptual grounds. In many cases, the designation is straightforward. For example, in our treatment study, it is theoretically natural to think of age as the moderator variable and to ask if the effects of the treatment generalize across different age groups. However, there also will be cases where the designation is not straightforward. For example, a study might examine depression as a function of gender and two levels of ethnicity, European American and African American, in the context of a 2×2 factorial design. In such a framework, one could treat gender as the focal independent variable and ask if gender differences change as a function of ethnicity. Or, one could treat ethnicity as the focal independent variable and ask if ethnic differences change as a function of gender. From a statistical standpoint, the designation of moderator variable status is arbitrary because the value of the relevant single degree of freedom interaction parameter will be the same in either designation. The choice depends on how the researcher wants to frame the discussion.

What about experimentwise controls for multiple contrasts? In the previous analysis, we conducted three single degree of freedom interaction contrasts. If one desires to control the experimentwise error rate across the contrasts, the modified Bonferroni method discussed in Appendix A can be used. For alternative control procedures that may be more powerful in some scenarios, see Westfall et al. (1999). As before, issues of statistical power should be of concern, and it is good practice to conduct the tests both with and without the experimentwise controls to evaluate the robustness of the conclusions across analytic method.

As noted previously, although simple main effects do not address features of statistical interaction, they are usually of conceptual interest. When a factor has more than a single degree of freedom, then a natural grouping of subcontrasts based on that factor is into three families defined by the single degree of freedom main effect contrasts, the single degree of freedom simple main effects contrasts, and the single degree of freedom interaction contrasts, respectively. These guidelines are arbitrary and can be overridden by theoretical or practical concerns. For additional discussion of this issue, see Toothaker (1993).

What role does the omnibus interaction effect have in the analysis? As before, the overall omnibus interaction effect usually will be of little conceptual interest. The exceptions would be if the experimentwise control procedure required the omnibus F test or if the researcher was interested in specifying the overall percent of variance accounted for by the interaction.

What about higher order interactions? In designs with three or more factors, higher order interactions also can be conceptualized in terms of single degree of freedom interaction parameter estimates. Appendix C describes such an example. See also Jaccard (1998).

Action Steps

When analyzing an interaction effect, it is important to evaluate the interaction using analytic methods that map onto the concept of interaction. Simple main effects analysis does not do so. Although such analyses will typically be informative, single degree of freedom interaction contrasts typically will be most revealing. Such analyses should be pursued using the principles for controlling experimentwise error rates discussed earlier. Specifically, the contrasts should be evaluated both with and without the experimentwise controls to determine if conclusions are robust across the two forms of analysis.

Journal Presentation

In this section, we characterize strategies for presenting the results of a factorial ANOVA when such analyses represent a major focus of a journal report. We focus on concepts discussed in this article. For additional issues that should be addressed explicitly in a journal presentation, see Jaccard and Guilamo-Ramos (in press).

Although they may not be of direct conceptual interest, it probably is useful to present the results of the omnibus F tests to provide readers a sense of the behavior of the overall factors and to permit readers who have “omnibus-like” interests to calculate their own statistics of interest. Because of space constraints, it is rare for researchers to report ANOVA summary tables focused on the omnibus F tests. If this is not done, then the text description of the omnibus tests should include the F ratio for the effect, the degrees of freedom, the p value, and the mean square error that was used in the calculation of the F ratio. The reporting of the mean square error permits researchers who prefer to evaluate indexes of effect size (e.g., eta squared and other such indexes) to compute them from the information.

Factorial ANOVAs typically will involve exploration of many single degree of freedom contrasts. The results for these contrasts are best reported in a table, especially if many of them are statistically significant. Table 3 presents a format using the 3 × 2 treatment study that can be used as a blueprint. The table contains the theoretically interesting single degree of freedom contrasts associated with the main effects, followed by the simple main effects and the interaction contrasts. The reported results for each of these contrasts include the estimated parameter value (e.g., the mean difference in question, the interaction parameter in question); the estimated standard error of the parameter (to provide a sense of sampling error); the *t* test evaluating the statistical significance of the parameter (sometimes this is reported as an *F* test, which will simply be *t*²); the degrees of freedom associated with the error term for the contrast; the *p* value for the significance test; and the 95% confidence interval. Reporting the estimated standard error of the contrast is important because this allows researchers to calculate additional informative statistics, as discussed in Jaccard and Guilamo-Ramos (in press). Reporting the *p* values is important because this permits researchers to invoke different experimentwise control procedures based on different definitions of families of contrasts. The table should be accompanied by a discussion of how families were defined for purposes of invoking experimentwise controls and what the results were when the controls were applied.

As an example, we provide a brief presentation of the results for the hypothetical treatment study as it might appear in a journal. We omit many important facets of the analysis that are discussed in Jaccard and Guilamo-Ramos (in press). Rather, our focus is on presentational matters for the concepts developed in this article. The ensuing discussion assumes that a table of cell means has been presented (which we refer to

as Table A) as well as Table 3 (which we refer to as Table B). The write-up might appear as follows:

The five outcome variables were clustered into two families consisting of the primary outcome measure (depression) and the secondary outcome measures (anxiety, fear reactions, quality of peer relations, and quality of family relations). A 3 × 2 between-subjects factorial ANOVA was performed on each of the five outcomes both with and without a Holm-based modified Bonferroni correction to control Type I error rates across outcomes for a given factor. None of the effects for the secondary outcomes were statistically significant, and this was true both with and without the application of the Holm procedure. For the depression outcome, both main effects and the interaction effect were statistically significant. The omnibus *F* for the treatment group main effect was 26.62 (*df* = 2, 294, *p* < 0.001), for the age main effect it was 71.44 (*df* = 1, 294, *p* < 0.001), and for the interaction effect between the two factors it was 19.73 (*df* = 2, 294, *p* < 0.001). The mean square error for each of these *F* ratios was 0.915. Table A presents the cell means, standard deviations, and sample sizes. Table B presents relevant single degree of freedom contrasts and their associated statistics.

Of primary theoretical interest was contrasts comparing (a) the cognitive treatment group to the control group, (b) the behavioral treatment group to the control group, and (c) the cognitive treatment group to the behavioral treatment group. The first three rows of Table B present these three comparisons for main effect means collapsing across age. Tests of statistical significance of these comparisons were performed both with and without experimentwise controls across the three contrasts (using the somewhat conservative Holm test), and the conclusions were comparable in both cases. The mean depression score for the cognitive treatment group was statistically significantly less than that for

Table 3. Example Tabular Format for Single Degree of Freedom Contrasts

	Parameter	Standard Error	<i>t</i> Value	<i>p</i> Value	Lower Limit	Upper Limit
ME: Cog-Cntrl	-.980	.135	7.26	<.001	-1.246	-.714
ME: Beh-Cntrl	-.390	.135	2.89	<.004	-.656	-.124
ME: Cog-Beh	-.590	.135	4.37	<.001	-.856	-.324
ME: Younger-Older	.933	.110	8.48	<.001	.716	1.151
SME: Cog-Cntrl for Younger	-1.820	.191	9.53	<.001	-2.196	-1.444
SME: Beh-Cntrl for Younger	-.700	.191	3.66	<.001	-1.076	-.324
SME: Cog-Beh for Younger	-1.120	.191	5.86	<.001	-1.406	-.744
SME: Cog-Cntrl for Older	-.140	.191	0.73	.465	-.516	.236
SME: Beh-Cntrl for Older	-.080	.191	0.42	.676	-.456	.296
SME: Cog-Beh for Older	-.060	.191	0.31	.754	-.436	.316
IC1: Cog, Cntrl by Younger, Older	-1.680	.271	6.23	<.001	-2.211	-1.151
IC2: Beh, Cntrl by Younger, Older	-.620	.271	2.30	.020	-1.152	-.092
IC3: Cog, Beh by Younger, Older	-1.060	.271	3.93	<.001	-1.592	-.533

Notes: ME = main effect contrast; SME = simple main effect contrast; IC = interaction contrast; Cog = cognitive treatment; Beh = behavioral treatment; Cntrl = control group; Younger = younger children; Older = older children. For interaction contrasts; IC1 = Cog-Cntrl for younger minus Cog-Cntrl for older; IC2 = Beh-Cntrl for younger minus Beh-Cntrl for older; IC3 = Cog-Beh for younger minus Cog-Beh for older; Lower limit and upper limit are for 95% confidence intervals. Degrees of freedom for the *t* tests = 294.

the control group (mean difference = $-.98$) as well as the behavioral treatment group (mean difference = $-.590$). In addition, the mean depression score for the behavioral treatment group was statistically significantly lower than that of the control group (mean difference = $-.590$). To determine if such differences occurred for each age group, simple main effect contrasts were performed. Rows 4 to 6 of Table B present statistics for the three comparisons for just the younger children, and rows 7 to 9 present statistics for the comparisons for just the older children. For the younger children, all three mean differences were statistically significant and followed the same general pattern as the main effects; the mean depression score for the cognitive treatment group was less than that of the control group (mean difference = -1.82) as well as the behavioral treatment group (mean difference = -1.120); the mean depression score for the behavioral treatment group was less than that of the control group (mean difference = $-.700$). These differences maintained statistical significance when a Holm test was applied across the three simple main effect contrasts to control the experimentwise error rate. For the older children, none of the comparisons were statistically significant.

To formally test if the group differences were statistically significantly stronger in younger as compared with older children, three single degree of freedom interaction contrasts were evaluated. The first compared the cognitive treatment group and the control group mean difference for younger children with that of older children. The second compared the behavioral treatment group and the control group mean difference for younger children with that of older children. The third compared the cognitive treatment group and the behavioral treatment group mean difference for younger children with that of older children. The last three rows of Table B present the relevant interaction parameters and their significance tests. All three tests were statistically significant, and this was true both with and without the application of the Holm procedure for controlling experimentwise error across the three contrasts. The mean difference for the cognitive treatment and control group for younger children ($1.98 - 3.80 = -1.82$) was statistically significantly different from the corresponding mean difference for older children ($1.96 - 2.10 = -.14$). This differential effect is reflected in the interaction parameter $(-1.82) - (-.14) = -1.68$. The mean difference for the behavioral treatment and control groups for younger children ($3.10 - 3.80 = -0.70$) was statistically significantly different from the corresponding mean difference for older children ($2.02 - 2.10 = -0.08$). This differential effect is reflected in the interaction parameter $(-0.70) - (-.08) = -0.62$. Finally, the mean difference for the cognitive treatment and behavioral treatment groups for younger children ($1.98 - 3.10 = -1.12$) was statistically significantly different from the corresponding mean difference for older children ($1.96 - 2.02 = -.06$).

This differential effect is reflected in the interaction parameter $(-1.12) - (-.06) = -1.06$.

Computer Programs

This article has emphasized the importance of single degree of freedom contrasts focused on main effects, simple main effects, and interactions. All of the major computer statistical packages permit the calculation of such contrasts, although doing so is sometimes cumbersome and involved. In SPSS, single degree of freedom contrasts are calculated using the general linear model program in conjunction with the /LMATRIX subcommand. Alternatively, they can be isolated by requesting the parameter estimates options. Interpretation of the parameter estimates using this option requires knowledge of the dynamics of dummy variables in linear models (Jaccard, 1998). The Web site www.zumastat.com provides a set of programs that can be integrated with SPSS or that can serve as stand-alone programs to easily calculate parameter estimates, their standard errors, significance tests, and confidence intervals for a wide range of factorial designs. The program uses as input the cell means, cell sample sizes, and the overall error term as derived from another statistical package. The programs can be used for designs with repeated measures as well as analysis of covariance.

References

- Bernhardson, C. S. (1975). Type I error rates when multiple comparison procedures follow a significant F test of ANOVA. *Biometrics*, *31*, 229–232.
- Boik, R. J. (1979). Interactions, partial interactions, and interaction contrasts in the analysis of variance. *Psychological Bulletin*, *86*, 1084–1089.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.
- Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I error and power of latent mean methods and MANOVA in factorially invariant and non-invariant latent variable systems. *Structural Equation Modeling*, *7*, 534–556.
- Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association*, *1*, 1000–1004.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*, 800–802.
- Holland, B. S., & Copenhaver, M. (1988). Improved Bonferroni-type multiple testing procedures. *Psychological Bulletin*, *104*, 145–149.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.
- Holmbeck, G. (1997). Toward terminological, conceptual, and statistical clarity in the study of mediators and moderators: Examples from the child-clinical and pediatric psychology litera-

- tures. *Journal of Consulting and Clinical Psychology*, 65, 599–610.
- Jaccard, J. (1998). *Interaction effects in factorial analysis of variance*. Newbury Park, CA: Sage.
- Jaccard, J., & Guilamo-Ramos, V. (in press). Analysis of variance frameworks in clinical child and adolescent psychology: Advanced issues and recommendations. *Journal of Clinical Child and Adolescent Psychology*.
- Kirk, R. (1995). *Experimental design: Procedures for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole.
- Morrison, D. (1959). *Multivariate statistics*. New York: McGraw-Hill.
- Seaman, M. A., Levin, K. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin*, 110, 577–586.
- Toothaker, L. E. (1993). *Multiple comparison procedures*. Newbury Park, CA: Sage.
- Westfall, P., Tobias, R., Rom, D., Wolfinger, R., & Hochberg, Y. (1999). *Multiple comparisons and multiple tests: Using the SAS system*. Cary, NC: SAS Institute.
- Wilcox, R. R. (1996). *Statistics for the social sciences*. New York: Academic.
- Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. New York: Academic.
- Wilcox, R. R. (2001). *Fundamentals of modern statistical methods*. New York: Springer-Verlag.

Appendix A

Modified Bonferroni Methods

Holm (1979; see also Holland & Copenhaver, 1988; Seaman, Levin, & Serlin, 1991) has suggested a modified Bonferroni method that is more powerful than the traditional Bonferroni-based approach but that adequately maintains experimentwise error rates at the desired alpha level. We first describe the approach conceptually and then give a concrete application for the 3×2 intervention study.

Conceptual Description

First, a p value is obtained for each contrast in the family of contrasts. The p values are then ordered from smallest to largest. If two p values are identical, they are ordered arbitrarily or using theoretical criteria. The contrast with the smallest p value is evaluated against an alpha of $.05/k$, where k is the total number of contrasts in the family. If this leads to rejection of the corresponding null hypothesis (because the observed p value is less than the adjusted α), then the next smallest p value is tested against an alpha level of $.05/(k - 1)$, where $k - 1$ is the remaining number of contrasts. If this test leads to null hypothesis rejection, then the next smallest p value is tested against an alpha level of $.05/(k - 2)$, and so on until a nonsignificant difference is observed. Once a statistically nonsignificant difference is observed, all remaining contrasts are declared nonsignificant.

Example Application

In our example, there were five outcome variables, so we conduct five separate univariate ANOVAs. Suppose we focus on the main effect of the treatment factor. We first find the analysis of the five that yielded the smallest p value. If this p value is less than $.05/5 = .01$, then we conclude that the effect is statistically significant. We next choose the analysis that had the second smallest p value. If this p value is less than $.05/4 = .0125$, then we conclude that the effect for this outcome is statistically significant. We then find the analysis with the third smallest p value and contrast it with a critical alpha of $.05/3 = .0167$. The analysis with the fourth smallest p value is compared against a critical alpha level of $.05/2 = .025$, and the analysis with the largest observed p value is compared with an alpha level of $.05/1 = .05$. If at any point in this sequential process an effect is declared to be statistically nonsignificant, then all remaining effects (i.e., all effects with larger p values) are declared as statistically nonsignificant as well.

Additional Bonferroni Variants

The Holm method is a “step-down” method in that one adjusts the critical value for the smallest p value, then the second smallest, and so on until the largest one is reached and evaluated against an alpha level of $.05$. An alternative approach is to use a “step-up” procedure, such as that suggested by Hochberg (1988). The Hochberg approach is identical to the Holm procedure, but works in the reverse direction, from the largest p value to the smallest. If the largest p value in the family of contrasts is less than $.05$, then all contrasts are declared statistically significant. If the largest p value is greater than $.05$, but the next largest one is less than $.05/4$ (using our example where there are five contrasts in the family), then the contrast in question and all those with smaller p values are declared statistically significant, and so on. The Hochberg method has slightly more statistical power than the Holm method and hence may be preferable. However, it does not control experimentwise error rates as well as the Holm method under some error structures. For more details, see Westfall et al. (1999).

Appendix B

Specification of Simple Main Effect and Interaction Contrasts

We illustrate the approach for calculating single degree of freedom contrasts using the 3×2 example, which is a completely between-subjects design. There are four steps for testing a single degree of freedom

contrast: (a) calculation of the parameter estimate, (b) calculation of the estimated standard error of the parameter, (c) calculation of the *t* test, and (d) calculation of confidence intervals.

Calculation of Parameter Estimates

As a first step, we must identify the contrast that is of theoretical interest and then calculate the value of the parameter estimate for that contrast. Suppose the investigator wanted to know if the mean score for the cognitive group was statistically significantly higher than the mean score for the control group when only younger children were considered. This corresponds to a simple main effect, because we are examining the effect of the independent variable on the dependent variable at a single level of the moderator variable. From Table 2, we see that the mean for the cognitive treatment group for younger children was 1.98 and the mean for the control group for younger children was 3.80. The difference between these two means is the parameter value for this simple main effect, and it equals $1.98 - 3.80 = -1.82$. We want to test the significance of this difference.

There is another way of isolating this parameter value of -1.82 that uses weighting coefficients. The approach appears cumbersome, but it greatly simplifies later computations. We begin by listing each cell mean for the factorial design on the right side of an equation with a weighting coefficient attached to each mean:

$$PE = c_1 M_{Cog,Young} + c_2 M_{Beh,Young} + c_3 M_{Cntl,Young} + c_4 M_{Cog,Old} + c_5 M_{Beh,Old} + c_6 M_{Cntl,Old} \quad (1)$$

where *PE* stands for the parameter estimate that we want to compute, the various *M* indicate cell means, and the various *c* are weighting coefficients. We want to define values of *c* so that we produce $M_{Cog,Young} - M_{Cntl,Young} = 1.98 - 3.80 = -1.82$. If we substitute all of the mean scores from Table 2 into Equation 1, we obtain

$$PE = c_1 1.98 + c_2 3.10 + c_3 3.80 + c_4 1.96 + c_5 2.02 + c_6 2.10 \quad (2)$$

The key is to assign values to the *c* coefficients that isolate the parameter estimate of interest. On some reflection, it can be seen that if we set $c_1 = 1$, $c_3 = -1$, and all other coefficients to zero, we get:

$$PE = 1 M_{Cog,Young} + 0 M_{Beh,Young} + -1 M_{Cntl,Young} + 0 M_{Cog,Old} + 0 M_{Beh,Old} + 0 M_{Cntl,Old}$$

Carrying out the multiplication reduces this to

$$PE = 1 M_{Cog,Young} + -1 M_{Cntl,Young} = M_{Cog,Young} - M_{Cntl,Young}$$

which is the mean difference we are interested in. Using Equation 2, we get

$$PE = (1)(1.98) + (0)(3.10) + (-1)(3.80) + (0)(1.96) + (0)(2.02) + (0)(2.10) = 1.98 - 3.80 = -1.82$$

The reason that we prefer this more cumbersome method to calculating the mean difference is because the values of *c* are necessary later to carry out the significance test.

As another example, suppose that we were theoretically interested in comparing the mean for the cognitive therapy group minus the mean for the control group but for only the older children. Using Equation 2 and on some reflection, it can be seen that this is accomplished by setting $c_4 = 1$, $c_6 = -1$, and all other *c* values to zero:

$$PE = 0 M_{Cog,Young} + 0 M_{Beh,Young} + 0 M_{Cntl,Young} + 1 M_{Cog,Old} + 0 M_{Beh,Old} + -1 M_{Cntl,Old} = (0)(1.98) + (0)(3.10) + (0)(3.80) + (1)(1.96) + (0)(2.02) + (-1)(2.10) = 1.96 - 2.10 = -0.14$$

Any parameter estimate in Table 3 can be calculated by assigning appropriate values to the *c* coefficients. We now describe the *c* values that one would assign to isolate the parameter estimates for each entry in Table 3.

The first three entries are single degree of freedom contrasts for the main effect means for the treatment factor. Each comparison requires that we average the means across the age factor. For the first entry in the table, we want to compare the average of the two cognitive therapy means, $(M_{Cog,Young} + M_{Cog,Old})/2$, with the average of the two control group means, $(M_{Cntl,Young} + M_{Cntl,Old})/2$. The average of the two cognitive therapy means, $(M_{Cog,Young} + M_{Cog,Old})/2$, is equivalent to the expression $0.5 M_{Cog,Young} + 0.5 M_{Cog,Old}$, so this average can be isolated by assigning the value of 0.5 to c_1 and c_4 . Similarly, the average of the two control group means is equivalent to the expression $0.5 M_{Cntl,Young} + 0.5 M_{Cntl,Old}$, so this can be isolated by assigning the value of 0.5 to c_3 and c_6 . Because we want the parameter estimate to equal the *difference* between the two sets of averaged means, we set $c_1 = 0.5$, $c_4 = 0.5$ and $c_3 = -0.5$ and $c_6 = -0.5$. Note that we obtain this difference by reversing the initial signs we gave to c_3 and c_6 . This yields

$$PE = .5 M_{Cog,Young} + 0 M_{Beh,Young} + -.5 M_{Cntl,Young} + .5 M_{Cog,Old} + 0 M_{Beh,Old} + -.5 M_{Cntl,Old} = (.5)(1.98) + (0)(3.10) + (-.5)(3.80) + (.5)(1.96) + (0)(2.02) + (-.5)(2.10) = -0.98$$

Verify from the data in Table 2 that this yields the value of interest. Using similar logic, the coefficients for all the main effect parameter values in Table 3 are

	c_1	c_2	c_3	c_4	c_5	c_6
ME: Cog – Cntrl	.5	0	–.5	.5	0	–.5
ME: Beh – Cntrl	0	.5	–.5	0	.5	–.5
ME: Cog – Beh	.5	–.5	0	.5	–.5	0
ME: Younger – Older	.33	.33	.33	–.33	–.33	–.33

The next six entries in Table 3 correspond to simple main effect single degree of freedom contrasts. The two examples we used previously to first introduce how to use Equation 1 isolated simple main effects. For this reason, we will not repeat the exposition of these. The coefficients for the six contrasts are

	c_1	c_2	c_3	c_4	c_5	c_6
SME Cog – Cntrl for Younger	1	0	–1	0	0	0
SME Beh – Cntrl for Younger	0	1	–1	0	0	0
SME Cog – Beh for Younger	1	–1	0	0	0	0
SME Cog – Cntrl for Older	0	0	0	1	0	–1
SME Beh – Cntrl for Older	0	0	0	0	1	–1
SME Cog – Beh for Older	0	0	0	1	–1	0

Finally, the last three entries of Table 3 are single degree of freedom interaction contrasts. Consider the first of these entries. The 2×2 subtable that this interaction contrast focuses on is

	Younger	Older
Cog	1.98	1.96
Control	3.80	2.10

$$IC1 = (1.98 - 3.80) - (1.96 - 2.10) = -1.68$$

The difference between the cognitive therapy group and the control group for the younger children is isolated by setting $c_1 = 1$ and $c_3 = -1$. The difference between the cognitive therapy group and the control group for the older children is isolated by setting $c_4 = 1$ and $c_6 = -1$. Because we want to find the *difference* between these two differences, we change the signs of c_4 and c_6 . This yields:

$$\begin{aligned}
 PE &= 1 M_{Cog,Young} + 0 M_{Beh,Young} + -1 M_{Cntl,Young} + \\
 &\quad -1 M_{Cog,Old} + 0 M_{Beh,Old} + 1 M_{Cntl,Old} \\
 &= (1)(1.98) + (0)(3.10) + (-1)(3.80) + (-1)(1.96) + \\
 &\quad (0)(2.02) + (1)(2.10) \\
 &= -1.68
 \end{aligned}$$

Using similar logic, the coefficients for the three interaction contrasts in Table 3 are

	c_1	c_2	c_3	c_4	c_5	c_6
IC1: Cog, Cntrl by Younger, Older	1	0	–1	–1	0	1
IC2 Beh, Cntrl by Younger, Older	0	1	–1	0	–1	1
IC3 Cog, Beh by Younger, Older	1	–1	0	–1	1	0

Calculation of the Estimated Standard Error

Once the value of the parameter estimate is obtained, the next step is to calculate the estimated standard error for the parameter. This is based on the cell sample sizes, the values of the c coefficients that were used when deriving the parameter estimate, and the mean square error from the analysis of variance summary table. The formula is

$$SE = \sqrt{MS_{ERROR} \left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \frac{c_3^2}{n_3} + \frac{c_4^2}{n_4} + \frac{c_5^2}{n_5} + \frac{c_6^2}{n_6} \right)}$$

where n is the sample size for the cell that corresponds to the associated c coefficient, and MS_{ERROR} is the mean square error from the overall analysis of variance summary table. As an example, in our treatment study, each cell mean was based on a sample size of 50. The MS_{ERROR} from the overall summary table was 0.91. For the first interaction contrast in Table 3 (IC1), the c coefficients were $c_1 = 1$, $c_2 = 0$, $c_3 = -1$, $c_4 = -1$, $c_5 = 0$, $c_6 = 1$, which yields

$$SE = \sqrt{0.91 \left(\frac{1^2}{50} + \frac{0^2}{50} + \frac{-1^2}{50} + \frac{-1^2}{50} + \frac{0^2}{50} + \frac{1^2}{50} \right)} = 0.27$$

The estimated standard error of the first interaction contrast is 0.27.

t Test

The t ratio for the test of significance of the parameter estimate is the absolute value of the parameter estimate divided by its estimated standard error, PE/SE . For the first interaction contrast in Table 3 (IC1) this equals the absolute value of $-1.68/0.27$, which is 6.23. The degrees of freedom for the t test is the degrees of freedom associated with the MS_{ERROR} from the overall summary table.

Confidence Intervals

The confidence intervals for the parameter estimate are calculated using the classic formula for confidence intervals. The lower limit is $PE - SE (t_{crit})$ and the up-

per limit is $PE + SE(t_{crit})$, where t_{crit} is the critical value of t used to determine a statistically significant result in the t test.

Technical Comments

The approach described here is applied with the constraint that the sum of the cs must equal 0. It can be used with equal or unequal cell sample sizes. For unequal sample sizes, the resulting parameter estimate and standard error is identical to that obtained using standard least squares analysis of variance with a Type III sums of squares approach, which is the traditional solution used in psychology. Other modeling approaches require a more complex contrast structure. The formulation described previously can be applied to analysis of covariance designs, but adjusted means are the focus of analysis, and a correction factor for the covariates must be included in the estimation of the standard error of the contrast. For details, see Kirk (1995). For use of the strategy with designs that include repeated measures, see Jaccard and Guilamo-Ramos (in press).

Appendix C

Example of a Three-Way, Single Degree of Freedom Interaction Parameter

We illustrate a single degree of freedom three-way interaction contrast using a $2 \times 2 \times 2$ between-subjects design. The outcome variable is a measure of self esteem that ranges from 0 to 80. Higher scores indicate higher levels of self esteem. The first factor is a treatment condition (treatment to raise self esteem vs. control group), the second factor is gender (girl vs. boy), and the third factor is ethnicity (African American vs. European American). Following the framework of Jaccard (1998), the first step is to specify a focal independent variable whose effects on the outcome variable you want to make the primary focus of analysis. In this case, the focal independent variable is chosen by the investigator to be the treatment condition. The second step is to specify a first-order moderator variable that is hypothesized to moderate the impact of the focal independent variable on the outcome variable. The investigator selects gender as the first-order moderator variable and, based on past literature, hypothesizes that the treatment minus control mean difference will be larger for girls than boys. The third step is to specify a second-order moderator variable. A second-order moderator variable moderates the impact of the first-order

moderator on the relationship between the focal independent variable and the dependent variable. Ethnicity is chosen as the second-order moderator. The investigator wants to determine if the interactive relationship between gender and treatment condition is different for African Americans as compared with European Americans. The mean values for the eight groups can be organized in the form of two 2×2 subtables involving the first-order moderator and the focal independent variable:

Subtable 1 (African American)			Subtable 2 (European American)		
	Female	Male		Female	Male
Treatment	76.25	30.35	Treatment	57.30	31.40
Control	42.55	29.90	Control	36.75	27.35
$(76.25 - 42.55) - (30.35 - 29.90) = 33.25$			$(57.30 - 36.75) - (31.40 - 27.35) = 16.50$		

In the first subtable, the two-way interaction is characterized by the difference between column mean differences as a function of the first-order moderator variable, and the two-way interaction parameter estimate (which appears directly beneath the table) is 33.25. This is the sample parameter estimate for the two-way interaction at the first level of the second-order moderator. The same calculations characterize the two-way interaction for the second subtable, 16.50. This is the sample parameter estimate for the two-way interaction at the second level of the second-order moderator. If the two-way interactions are identical for both levels of the second-order moderator, then the difference between the two, two-way interaction parameter values should be zero. But in this case, the difference is non-zero (i.e., $33.25 - 16.50 = 16.75$), suggesting that a three-way interaction effect exists. These data represent samples, and it is possible that the three-way interaction parameter value in the population is zero and that the observed nonzero result of 16.75 is due to sampling error. The F ratio for the single degree of freedom three-way interaction contrast tests the viability of such an interpretation. A statistically significant effect indicates that the probability that the obtained result would occur if the null hypothesis of a zero three-way interaction parameter value is true is highly unlikely. For designs where some of the factors have more than a single degree of freedom, there will be more than one single degree of freedom three-way interaction contrast parameter (as was the case for the different two-way interaction parameters in Table 2). Each is defined using relevant subtables, using principles similar to those in Table 2 (see Jaccard, 1998).