Journal of Clinical Child and Adolescent Psychology
2003, Vol. 32, No. 3, 467–486

METHODOLOGICAL ARTICLE

Analytic Methods for Questions Pertaining to a Randomized Pretest, Posttest, Follow-Up Design

Joseph R. Rausch, Scott E. Maxwell, and Ken Kelley
Department of Psychology, University of Notre Dame

Delineates 5 questions regarding group differences that are likely to be of interest to researchers within the framework of a randomized pretest, posttest, follow-up (PPF) design. These 5 questions are examined from a methodological perspective by comparing and discussing analysis of variance (ANOVA) and analysis of covariance (ANCOVA) methods and briefly discussing hierarchical linear modeling (HLM) for these questions. This article demonstrates that the pretest should be utilized as a covariate in the model rather than as a level of the time factor or as part of the dependent variable within the analysis of group differences. It is also demonstrated that how the posttest and the follow-up are utilized in the analysis of group differences is determined by the specific question asked by the researcher.

The randomized pretest, posttest, follow-up (PPF) design is a common experimental design for testing hypotheses about intervention effects in clinical child and adolescent research. In PPF designs, some outcome variable (e.g., depression, self-esteem) is measured on three separate occasions: once prior to the initiation of the treatment, once at the conclusion of the treatment, and once a specified time period after the conclusion of the treatment. For example, a researcher may record a depression score before the treatment is implemented, randomly assign participants to groups, record a second depression score at the conclusion of the treatment, and finally collect a third measurement 6 months after the conclusion of the treatment. This design is generally implemented for the purpose of determining if a treatment effect exists at the conclusion of the treatment and persists for some specified period of time after the treatment has ended. The randomized PPF design can also be utilized for answering a variety of questions about change over time when making group comparisons.

Although the randomized PPF design is a relatively common design in clinical research, confusion often exists among researchers with respect to the questions that can be asked and the appropriate methods for answering these questions. A number of analytic methods are possible when testing hypotheses about treatment effects within the context of a randomized PPF design such as analysis of variance (ANOVA), analysis of covariance (ANCOVA), and hierarchical linear modeling (HLM). Depending on the question(s) of interest, any of these and other analytic techniques may be possible, oftentimes leaving researchers perplexed when attempting to choose the most appropriate analytic technique for a specific question.

This article delineates the similarities and differences among various analytic techniques for the following five questions in a randomized PPF design that are likely to be of interest to researchers: (a) "Do the groups differ in any way over time?" (b) "Do the groups differ in change from the pretest to the posttest?" (c) "Do the groups differ in change from the pretest to the follow-up?" (d) "Do the groups differ in change from the posttest to the follow-up?" and (e) "Do the groups differ on the average of the posttest and the follow-up?" Although several analytic methods may plausibly answer a particular question of interest, within the context of this article the most appropriate method of analysis also provides an unbiased estimate of the parameter associated with the question of interest. Further, of the methods that provide an unbiased estimate, the most appropriate analytic method also provides the most statistical power and precision.

We compare the relative statistical power and precision of ANOVA and ANCOVA for these five questions about group differences over time in derivations found in Appendixes A and B. We also provide an illustrative

We would like to thank Stacey S. Poponak for her valuable comments on previous drafts of this article.

Requests for reprints should be sent to Joseph R. Rausch, Department of Psychology, University of Notre Dame, 118 Haggar Hall, Notre Dame, IN 46556. Email: jrausch@nd.edu

example along with significance tests and confidence intervals that are used to demonstrate the conceptual points for these five questions about group differences over time developed throughout the article. Further, we briefly discuss how HLM compares to ANCOVA when answering these five questions of interest about group differences within the context of a randomized PPF design.

Randomized Pre–Post Design

The discussion of analytic method comparisons begins within the context of the randomized pre–post design because the PPF design is an extension of the pre–post design. Questions asked within the PPF framework that include the pretest and either the posttest alone or the follow-up alone can be thought of as pertaining solely to a pre–post design from a statistical perspective. Thus, analytic methods that answer the question regarding group differences from the pretest to the posttest within the context of a randomized pre–post design can also be used when asking questions about group differences in change from the pretest to the posttest or the pretest to the follow-up within a randomized PPF design.

This section covers the assumptions required for the hypotheses tested within the randomized pre–post design to be valid and also reviews past methodological work on this design. Analytic methods that have been proposed for the randomized pre–post design are a one-within, one-between ANOVA in which the within-subjects factor is time and the between-subjects factor is group status (i.e., experimental condition), an ANOVA on the difference score, an ANCOVA on the posttest utilizing the pretest as a covariate, and an ANCOVA on the difference score utilizing the pretest as a covariate.[1]

Assumptions Within the Context of a Randomized Pre–Post Design

To facilitate proper interpretation and strengthen internal validity within a pre–post design, we assume that participants are randomly assigned to groups throughout the article. Although this is not literally an assumption of the analytic methods compared here, random assignment is necessary to equate the groups on the covariate and all other concomitant variables in the long run, allowing causal inferences to be made about the treatment effect.

The following statistical assumptions underlie the hypothesis tests and confidence intervals for the group main effect and time by group interaction within the one-within, one-between ANOVA, the ANOVA on the difference score, and the ANCOVA on the posttest covarying the pretest within the context of a randomized pre–post design. These three assumptions are as follows: (a) the dependent variable (e.g., the posttest, the difference score) is normally distributed in the population within each group (and conditional on the observed pretest scores when using ANCOVA); (b) the scores of different participants are statistically independent of one another (e.g., at a particular time point, observations are independent of one another); and (c) the population variance of the dependent variable is equal for all the groups (i.e., homogeneity of variance).

For ANCOVA on the posttest covarying the pretest, we also assume homogeneity of regression slopes to simplify our presentation. Still, there may be situations where the homogeneity of regression assumption in ANCOVA is not tenable. If one suspects that the regression slopes for the treatment and control groups differ in the population, then it is necessary to explicitly add the pertinent parameter(s) to the statistical model. A researcher is then able to obtain a more complete understanding of the data by adding these parameters, which specify an interaction between the pretest scores and the treatment. Researchers interested in relaxing the homogeneity of regression assumption may consult Rogosa (1980), who provided a discussion of analytic methods dealing with nonparallel regression lines, and Huitema (1980, chapter 13), who explained a procedure known as the Johnson–Neyman technique to analyze nonparallel regression lines.

When utilizing ANCOVA in randomized designs, statistical power and precision depend on the population correlation between the dependent variable and the covariate (i.e., the pretest in the context of a randomized pre–post design), $\rho_{DV,COV}$. The larger the magnitude of $\rho_{DV,COV}$, the more statistical power and statistical precision ANCOVA will yield. All other things being equal, measurement error in the dependent variable or the covariate or both and failing to account for a nonlinear relationship between the dependent variable and the covariate both tend to decrease the magnitude of $\rho_{DV,COV}$.[2] However, when participants are

[1] Maxwell, O'Callaghan, and Delaney (1993) provided an introduction to ANCOVA that covers a variety of topics, some of which are not addressed in this article due to space limitations. Throughout this article, all covariance analyses utilize the pretest as the sole covariate unless otherwise stated. Although it is possible to collect and utilize more than one covariate within ANCOVA, we focus on incorporating the pretest as the sole covariate to simplify our presentation. For readers interested in ANCOVA with multiple covariates, Huitema (1980, chapter 8) provided a thorough discussion of this topic.

[2] The ANCOVA model typically utilized in practice only accounts for the linear relationship between the dependent variable and the covariate. Thus, not accounting for quadratic (cubic, quartic, and so on) relationships tends to decrease power and precision when compared to an analysis that does account for these relationships when they truly exist in the population.

randomly assigned to groups and the covariate is collected prior to the start of the treatment, the estimate of the treatment effect is still unbiased when either measurement error exists or one does not account for a nonlinear relationship. Thus, both these conditions tend to decrease power and precision, but neither bias the estimate of the treatment effect within the context of a randomized design.[3]

Also, the covariate (i.e., the pretest within the context of a randomized pre–post design) utilized within an ANCOVA should be measured before the initiation of the treatment to ensure statistical independence between the treatment and the covariate in the population (Maxwell & Delaney, 1990, pp. 382–384, case 3). If the covariate is measured after treatment has begun, the design is confounded because it is not known if the reason for any observed mean difference between the groups on the covariate is due to the treatment or sampling error. It is likely that this difference is due to treatment to some extent, and if this is the case, one will typically lose power by partialling some of the treatment variance out of the treatment effect when using the covariate measured after the initiation of the treatment. Thus, it is important to measure the covariate before the treatment is initiated when performing an ANCOVA for the purpose of increasing statistical power within a randomized study, allowing the researcher to obtain a more efficient answer to the question of interest.

Analysis of Data Collected From a Randomized Pre–Post Design

Now that the necessary assumptions have been stated, the analytic methods utilized within the context of a randomized pre–post design can be compared to determine which statistical method is the most appropriate. The first method we discuss conceptualizes the data in terms of a one-within, one-between ANOVA, in which time is a within-subjects factor and group status is a between-subjects factor. There are three possible omnibus tests that can be obtained from this framework: a time main effect, a group main effect, and a time by group interaction.

The time main effect answers the question "Averaging over the groups, are the pretest and the posttest different from one another?" This effect is generally not useful to researchers interested in group comparisons because this effect averages over the treatment and control groups, disregarding any possible differences between them. One might also be interested in the group main effect. The question that the group main effect attempts to answer is "Are the groups different on the average of the pretest (*Pre*) and the posttest (*Post*)?" This test does compare groups, and it does so by averaging the outcome variable over time for each group, making the sum, *Post* + *Pre*, the effective dependent variable for this test. Utilizing this dependent variable, the full model for the test of the group main effect can be expressed as

$$Post_{ij} + Pre_{ij} = \mu_{(Post+Pre)_j} + \varepsilon_{ij} \qquad (1)$$

where $\mu_{(Post+Pre)_j}$ denotes the population mean for group j ($j = 1, 2, \ldots, a$, where a is the total number of groups) on the dependent variable, *Post* + *Pre*, and ε_{ij} is the error for individual i ($i = 1, 2, \ldots, n_j$, where n_j is the sample size in group j) in group j. Equation 1 can be re-expressed in a more useful form for our purposes:

$$Post_{ij} = \mu_{Post_j} + (-1)(Pre_{ij} - \mu_{Pre}) + \varepsilon_{ij} \qquad (2)$$

where μ_{Pre} is the population grand mean on the pretest and μ_{Post_j} is the population mean on the posttest for group j.

Equation 2 explicitly demonstrates that the test of the group main effect restricts the regression slope predicting the posttest from the pretest to be –1. At the very least, we would not expect this restriction to be reasonable unless there is a negative correlation in the population between the pretest and the posttest.[4] Although a negative correlation between the pretest and the posttest is possible, in practice this situation is not likely. Further, even if the correlation between the pretest and the posttest is negative, this does not necessarily imply that we should restrict the regression slope predicting the posttest from the pretest to be –1. Thus, the test of the group main effect rarely provides the

[3]Although measurement error in the covariate does bias the estimate of the population regression coefficient predicting the dependent variable from the covariate, it does not bias the estimate of the treatment effect because group status is uncorrelated with the covariate in the population due to random assignment to groups (assuming the measurement error in the dependent variable and the covariate is uncorrelated along with other standard regression assumptions). Random assignment to groups also ensures that failing to account for a nonlinear relationship between the dependent variable and the covariate does not bias the estimate of the treatment effect, as the nonlinear component is uncorrelated with group status in the population.

[4]The reasoning underlying this statement is shown through the relationship

$$\beta_{Post,Pre} = \rho_{Post,Pre} \frac{\sigma_{Post}}{\sigma_{Pre}}$$

where $\beta_{Post,Pre}$ is the population regression slope predicting the posttest from the pretest, σ_{Post} is the population standard deviation of the posttest, σ_{Pre} is the population standard deviation of the pretest, and $\rho_{Post,Pre}$ is the population correlation between the posttest and the pretest. In practice, it is likely in most situations that the pretest and the posttest will be positively correlated, leading to the conclusion that the population regression coefficient predicting the posttest from the pretest is typically positive.

most powerful and precise answer to the researcher's question about group differences.

The final omnibus effect that can be tested within the one-within, one-between ANOVA is the time by group interaction. The test of this effect answers the question "Do the groups change differently from the pretest to the posttest?" or, equivalently within a randomized pre–post design, "Are the groups different at the posttest?" Although these questions are generally conceptually different from one another, within a randomized pre–post design, they are equivalent.

The reason for the equivalence between these two questions is random assignment to groups, which ensures the groups are equal on the mean of the pretest scores in the population, assuming the pretest is measured before the start of the treatment. Figure 1 helps to illustrate the equivalence of these two questions for two groups when there is a treatment effect in the population (left panel) and when there is not a treatment effect in the population (right panel). Notice that the left panel in Figure 1 illustrates that a treatment effect is present in the population because Group 2 demonstrates a mean change of 2 points from the pretest to the posttest, whereas Group 1 demonstrates no mean change from the pretest to the posttest. Thus, from the perspective of groups changing differently from the pretest to the posttest, the treatment effect is 2 points. Also, notice there is a 2-point difference between the groups at the posttest. Because the groups must be equal on the mean of the pretest in the population within the context of a randomized pre–post design, these two population quantities will always be identical, demonstrating that these two approaches are attempting to find the same population quantity.

Notice the graph on the right of Figure 1 may not appear to represent Group 2. However, the reason that this line is not visible is due to the mean trajectory for Group 1 lying directly on top of the mean trajectory representing Group 2. This situation represents a case in which there is no treatment effect in the population. Notice that in the right panel of Figure 1 we obtain a treatment effect of zero from both the comparisons of difference scores and the difference on the posttest perspectives. As in the left panel in Figure 1, the equivalence of the mean group differences obtained from the difference score and posttest score approaches is due to the groups being equal on the pretest in the population, and, within randomized studies, parallelism among the population group mean trajectories (i.e., equal population group mean difference scores) equates to no treatment effect on the posttest. Although both the left and right panels of Figure 1 depict a situation in which Group 2 remains constant over time, the general principle illustrated here is also true when both the groups change from the pretest to the posttest. Further, the same principle applies in situations in which more than two groups are included in the study.

Thus, the test of the time by group interaction within the context of a randomized pre–post design answers the questions "Do the groups change differently

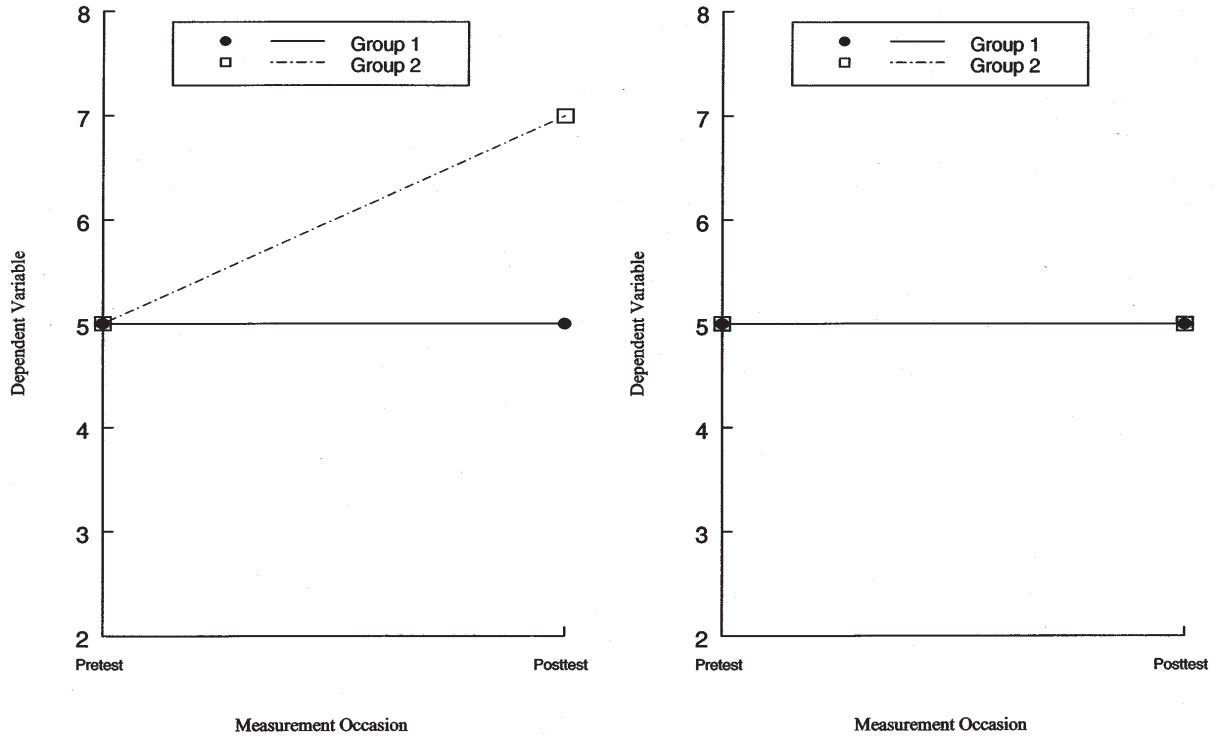

Figure 1. Plot of population group mean trajectories for two groups in which a treatment effect does (left panel) and does not (right panel) exist in the population.

from the pretest to the posttest?" and "Do the groups differ on the posttest?" As it has been shown these questions are equivalent in randomized designs. This statistical test utilizes *Post – Pre* as the effective dependent variable in the analysis when attempting to determine if the groups change differently from the pretest to the posttest. The full model for the test of the time by group interaction can be expressed in a manner similar to the test of the group main effect:

$$Post_{ij} - Pre_{ij} = \mu_{(Post-Pre)_j} + \varepsilon_{ij} \quad (3)$$

which can also be expressed as

$$Post_{ij} = \mu_{Post_j} + (1)(Pre_{ij} - \mu_{Pre}) + \varepsilon_{ij} \quad (4)$$

Equation 4 illustrates that the population regression slope predicting the posttest from the pretest is assumed to be 1 when testing the time by group interaction in a one-within, one-between ANOVA. This assumption is likely more reasonable in practice than the assumed slope of –1 for the test of the group main effect due to the positive correlation that is typically expected between the pretest and the posttest. Even though a positive correlation between the pretest and posttest implies the population regression slope predicting the posttest from the pretest will be positive, there is usually no reason to expect it to equal 1. Thus, restricting the population regression slope predicting the posttest from the pretest to be 1 generally leads to lower power and less precision than estimating this parameter from the data.

Another approach that has been popular in the literature is an ANOVA on the difference score, *Post – Pre*. Although an ANOVA on the difference score may seem to answer a different question than the time by group interaction, this analysis is mathematically equivalent to the interaction in the one-within, one-between ANOVA when analyzing data obtained from a pre–post design (Huck & McLean, 1975). In fact, one will receive identical observed *F* values and *p* values for these analyses for any data set from a pre–post design. Because of this, the shortcomings of the ANOVA on the difference score are the same as those encountered in the time by group interaction when testing for group differences in change within the context of a randomized pre–post design.

As Huck and McLean (1975) have shown, ANCOVA is generally the most appropriate analytic method when testing for group differences in change from the pretest to the posttest in a randomized pre–post design. It may seem that ANCOVA on the posttest covarying the pretest is only answering the question "Are the groups different on the posttest controlling for the pretest scores?" However, because the mean of the pretest scores for the different groups will be equal in the long run due to random assignment and the measurement of the pretest prior to the initiation of the treatment, the ANCOVA on the posttest is also answering the question "Do the groups change differently from the pretest to the posttest?" This result is one of the reasons why ANCOVA is useful when attempting to assess group differences in change within the context of randomized studies.

The advantage of the ANCOVA on the posttest can be seen from a more statistical perspective when comparing the following full model for ANCOVA to the full models for the group main effect and the time by group interaction within ANOVA that are illustrated in Equations 2 and 4, respectively:

$$Post_{ij} = \mu_{Post_j} + \beta_{Post,Pre}(Pre_{ij} - \mu_{Pre}) + \varepsilon_{ij} \quad (5)$$

As illustrated in Equation 5, ANCOVA allows for the data to estimate the population regression slope predicting the posttest from the pretest, $\beta_{Post,Pre}$, whereas the group main effect and the time by group interaction within the context of ANOVA implicitly constrain this value to be –1 and 1, respectively. Allowing $\beta_{Post,Pre}$ to be estimated from the data rather than restricting it to be –1 or 1 will generally reduce the population error variance in the model (at the expense of one denominator degree of freedom). Thus, the ANCOVA is generally a more powerful and precise procedure when compared to ANOVA when interest lies in group differences in change from the pretest to the posttest within the context of a randomized pre–post design. For example, suppose we utilize a randomized pre–post design for a two-group study in which the standardized group mean difference (i.e., δ, the population Cohen's *d*) is .5, the population correlation between the pretest and the posttest is .5, and the sample size for each group is 50. For this situation, assume that the population within-group variances of the pretest and the posttest are equal. When this is the case, the power for the ANCOVA on the posttest covarying the pretest is approximately .82, the power for both the ANOVA on the difference score and the ANOVA on the posttest alone is approximately .70, and the power for the group main effect in the one-within, one-between ANOVA is .30.

ANCOVA also controls for "unhappy randomization" (Kenny, 1979, p. 217), whereas the one-within, one-between ANOVA generally does not. Unhappy randomization occurs when random assignment produces groups that are significantly different on the pretest within a randomized pre–post design. Although this situation will not occur often (i.e., $100\alpha\%$ of the time, where α is the [unconditional] Type I error rate), inferences resulting from unhappy randomization can be flawed if one considers the conditional Type I error rate. The conditional Type I error rate is defined as the probability of falsely rejecting the null hypothesis of

group differences on the posttest given the observed values at the pretest. Within a randomized pre–post design, interest in the conditional Type I error rate corresponds with the question "Once pretest differences have been observed between the groups, can any differences at the posttest be trusted to reflect the treatment effect instead of the continuing influence of the difference at the pretest?" When random assignment is utilized and the pretest is measured before the start of the treatment, ANCOVA controls for these observed group differences on the mean of the pretest, controlling the conditional Type I error rate and allowing for valid inferences from this perspective.

Another analytic method that may seem reasonable is an ANCOVA on the difference score, *Post – Pre*, covarying the pretest because utilizing the difference score as the dependent variable explicitly answers the question of change, and utilizing the pretest as a covariate controls for any group differences on the mean of the pretest within the sample along with consistent individual differences from the pretest to the posttest. Although there is nothing necessarily wrong with this approach, this analysis is not necessary when testing for group differences in change from the pretest to the posttest within the context of a randomized pre–post design. The ANCOVA on the difference score covarying the pretest will yield the same F statistic and p value for the test of group differences as the ANCOVA on the posttest covarying the pretest for any particular data set. Thus, the only plausible reason for utilizing the ANCOVA on the difference score rather than the ANCOVA on the posttest is to facilitate the interpretation of change within each group (Hendrix, Carter, & Hintze, 1979).

Statistical Power Comparisons of ANOVA and ANCOVA Within the Context of a Randomized PPF Design

Now that we have determined that ANCOVA on the posttest covarying the pretest is the most appropriate analysis in a randomized pre–post design when interest lies in group differences in change from the pretest to the posttest or, equivalently, in group differences on the posttest, we demonstrate some general relations between ANOVA and ANCOVA. These relations will prove useful in determining which analytic method is the most appropriate in subsequent discussions about questions within randomized PPF designs. As in the randomized pre–post design, we assume the pretest is measured prior to the initiation of the treatment and participants have been randomly assigned to groups. Appendix A presents derivations of relevant standardized effect sizes for two or more groups. Appendix B shows how these standardized effect sizes compare to one another for various data analytic strategies. If two methods are identical and thus equivalent with respect to statistical power, they are also equivalent with respect to statistical precision. Similarly, if one method considered here is more statistically powerful than another method, it is also more precise. In this sense, the comparisons made in this section are for both statistical power and statistical precision.

Comparison of ANOVA and ANCOVA

When a researcher is contemplating the decision of analyzing data from a randomized PPF design with either ANOVA or ANCOVA, the results of Appendixes A and B will determine what analysis is the most appropriate choice. In particular, Appendix B demonstrates that regardless of the dependent variable that is analyzed from a randomized PPF design (e.g., *Post*, Follow-up [*Follow*], *Follow – Post*, *Post – Pre*, *Follow – Pre*), the pretest should almost always be used as a covariate. An ANCOVA that uses the pretest as a covariate will virtually always be more powerful than an ANOVA that utilizes the same dependent variable but ignores the pretest or an ANOVA that incorporates the pretest as a linear component of the dependent variable (e.g., an ANOVA on the difference score, *Follow – Pre*). Thus, whenever an ANOVA is performed when participants have been randomly assigned to groups and a pretest has been collected prior to treatment, it is virtually always a suboptimal analysis that will result in a loss of statistical power and precision when compared to the corresponding ANCOVA using the pretest as the covariate. For example, as stated within the previous section on the randomized pre–post design, one can think of the time by group interaction in the one-within, one-between ANOVA as utilizing *Post – Pre* as the dependent variable. Appendix B demonstrates that this analysis (as well as the group main effect) is suboptimal with respect to statistical power and precision when compared to ANCOVA on the posttest alone covarying the pretest.

Comparison of ANCOVAs

Although the previous section compared ANOVA to ANCOVA, it is also possible to compare different ANCOVAs to one another because the pretest is sometimes included as a linear component of the dependent variable. For example, one researcher might choose to utilize ANCOVA on the posttest alone using the pretest as the covariate, whereas another researcher might decide to use ANCOVA on the difference score, *Post – Pre*, using the pretest as the covariate. Appendix A demonstrates that the power and precision for ANCOVA on the difference score and ANCOVA on the posttest (both covarying the pretest) are equivalent. Further, the observed F values, p values, and confidence intervals for group mean comparisons will be equal for both of these

analyses for any given data set. Thus, not only are power and precision unaffected by utilizing the pretest as a component of the dependent variable in ANCOVA, but the statistical results in the sample associated with group mean comparisons are also unaffected.

We may also be interested in comparing ANCOVA on the posttest to ANCOVA on the dependent variable, *Post* + *Pre*. The fundamental message of this section is that these analyses will yield the same statistical power, and all ANCOVAs that covary the pretest and add (or subtract) some multiple of the pretest to the same dependent variable will not only yield the same statistical power and precision, but will also yield the same F value and p value for the test of the treatment effect along with the same confidence intervals for group mean comparisons for a particular data set. Thus, the only plausible reason for incorporating the pretest as part of the dependent variable and as a covariate in an ANCOVA is to facilitate the interpretation of the dependent variable within each group as was the case in the randomized pre–post design.

Randomized PPF Design

Realizing that ANCOVA using the pretest as a covariate generally provides more statistical power to detect treatment effects and more statistically precise confidence intervals around population group mean differences than ANOVA within the context of randomized designs, we utilize this methodological thinking within the context of the randomized PPF design. Because there are more questions of potential interest in a PPF design than in a pre–post design, there are also more methods that can potentially be chosen to analyze data from a randomized PPF design. We compare some plausible analytic methods for the five questions we believe are likely to be of the most interest to researchers working within the randomized PPF framework. It is important to remember the assumptions that were made for the randomized pre–post design because these same assumptions are also utilized for the randomized PPF design. The only difference is the analyzed dependent variable (e.g., *Follow* – *Post*, *Follow*) will change depending on the particular question to be answered.

Illustrative Example

To better illustrate the analytic perspectives described in the remainder of the article, a hypothetical data set is provided.[5] Suppose a researcher is interested in the effects of different forms of therapy intervention on childhood depression and plans to examine the effects of two therapies, Treatment A and Treatment B, and also includes a control group to eliminate alternative rival hypotheses (Campbell & Stanley, 1963). We will suppose that higher depression scores indicate a more depressed child.

After a pretest measure is collected from each child, 25 children are randomly assigned to each of the three groups without regard to their pretest scores.[6] After the therapy sessions have concluded, a posttest measure of depression is collected for each child. Six months after the collection of the posttest, a follow-up measure of depression is collected for each child to assess the lasting effects of therapy. Thus, each of the 25 children in the three groups has been assessed at three time points: before treatment, immediately following the conclusion of treatment, and 6 months after the conclusion of treatment. A graphical display of the mean trajectory for each group is provided in Figure 2, whereas Table 1 shows some relevant descriptive statistics for this hypothetical data set. Although other descriptive and inferential statistics, as well as a variety of other figures, may be of interest for a given research setting, the purpose of Figure 2 and Table 1 is to provide a context for the analytic results we discuss momentarily.

Figure 2 and Table 1 show that all three of the group means on the pretest are similar to one another. When randomly assigning to groups, this will typically be the case, and the group means on the pretest must be equal in the population. Within these sample data, the mean trajectory for the control group generally maintains its initial elevation over the three occasions of measurement, indicating no appreciable change in depression scores. The mean depression scores in Treatment A and Treatment B both decline from the pretest to the posttest, yet Treatment B appears to show a steeper decline within this time interval. From posttest to follow-up, the mean trend of Treatment A generally maintains the same level of depression. The mean trend of Treatment B declines to some extent from the posttest depression measure to the follow-up, illustrating Treatment A and Treatment B apparently differ in change within this time interval. Of course, inferential statisti-

[5]The data set along with the S-Plus or SPSS code used to generate the results is available by contacting the first author.

[6]Rather than simple random assignment to condition, another alternative is to form b blocks (the number of participants in each block is equal to N divided by b, where b is an arbitrary number chosen for the number of blocks and N is the total sample size) and randomly assign participants to groups within each block. The blocks are formed by placing participants with similar values on the pretest into a particular block. Although random assignment is guaranteed to equate the groups on the mean of the pretest in the long run, for any given study the groups will generally differ on their mean pretest scores. When random assignment to condition after blocking on the pretest is employed, the groups are generally more equivalent on the pretest measure than if the pretest is ignored within the random assignment procedure, increasing power while still maintaining internal validity. It is important to emphasize that blocking specifies an assignment procedure here and not a method of data analysis (see Matthews [2000] and Friedman, Furber, & DeMets [1998] for information on methods of randomly assigning participants to groups).

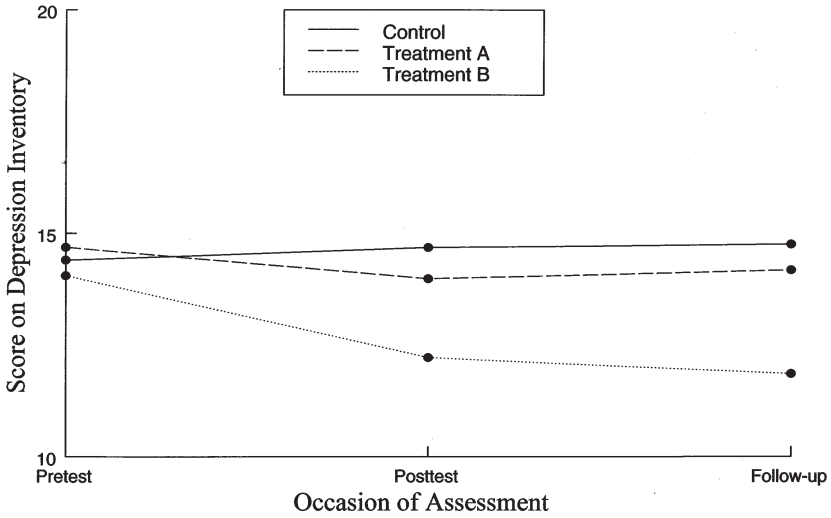

Figure 2. Plot of mean group trajectories for the illustrative data set.

Table 1. *Descriptive Statistics for Illustrative Data Set*

Pooled within-group correlation coefficients

Pretest	1		
Posttest	.4573	1	
Follow-up	.3872	.4164	1

Within-group means and standard deviations for each measurement occasion

	Pretest		Posttest		Follow-Up	
	M	*SD*	*M*	*SD*	*M*	*SD*
Control	14.394	3.771	14.679	3.234	14.756	3.516
Treatment A	14.682	4.527	13.975	4.189	14.172	5.254
Treatment B	14.045	3.246	12.208	3.367	11.849	3.173

cal methods (e.g., hypothesis tests, confidence intervals) are needed to assess which, if any, of the apparent effects in Figure 2 may simply reflect sampling error.

Five Questions of Interest Within the Context of a Randomized PPF Design

As stated in the introductory section, there are five substantive questions that may be of interest to researchers working within the context of randomized PPF designs. Each of these questions is delineated in the following subsections, along with the results from the hypothetical data set. Although there may be other potentially interesting questions that can be asked from PPF designs, the five questions we discuss are presumably the most useful for many research settings.

Group Differences in Any Way Over Time

The first question "Is there any evidence that the groups differ in any way over time?" is answered by comparing groups across time to probabilistically infer whether the groups differ on the outcome variable at any measured time point. We begin with this question because it subsumes all other patterns of group differences. Thus, its strength comes from its generality. However, it is so general that oftentimes researchers may decide to skip this question and proceed immediately to more specific questions and accompanying analyses. Nevertheless, we begin with this question largely because it establishes a conceptual framework for the remaining questions.

Analytic methods corresponding to the question of group differences in any way over time can simultaneously incorporate the posttest and the follow-up scores as dependent variables yielding a multivariate analysis. As was the case in univariate statistical tests delineated in previous sections of this article, the appropriate usage of the pretest measure is to treat it as a covariate. Thus, a multivariate analysis of covariance (MANCOVA) that uses the pretest as a covariate is generally more statistically powerful when testing for treatment effects than a multivariate analysis of vari-

Table 2. *MANOVA and MANCOVA Omnibus Tests for Various Sets of Dependent Variables*

	Post* and *Follow		***D* and *M***		***Post – Pre* and *Follow – Pre***	
Omnibus Test	**MANCOVA**	**MANOVA**	**MANCOVA**	**MANOVA**	**MANCOVA**	**MANOVA**
Observed *F* value	$F(4, 140) = 2.538$	$F(4, 142) = 2.250$	$F(4, 140) = 2.538$	$F(4, 142) = 2.250$	$F(4, 140) = 2.538$	$F(4, 142) = 1.311$
P value	$p = .043$	$p = .067$	$p = .043$	$p = .067$	$p = .043$	$p = .269$

Note: MANOVA = multivariate analysis of variance; MANCOVA = multivariate analysis of covariance. All *F* statistics and *p* values are based on the Wilks's lambda criterion.

ance (MANOVA) that ignores the pretest entirely or utilizes the pretest as a linear component of one or both of the dependent variables. Because of this, we recommend the MANCOVA approach when using a multivariate technique to answer the question regarding whether groups differ in any way over time.[7]

We delineate two potential approaches from the MANCOVA perspective. The first MANCOVA approach utilizes the posttest and the follow-up simultaneously as dependent variables and the pretest as a covariate. The second conceptualization of the MANCOVA approach assumes that a researcher might want to consider group differences from a different perspective. For example, a researcher might want to test whether groups differ in terms of an average of the posttest and the follow-up and a difference between follow-up and posttest. To answer these questions, one can utilize what we will define as *M* and *D* variables as the two dependent variables analyzed simultaneously while the pretest is used as a covariate. The average of posttest and follow-up, *M*, and the difference between follow-up and posttest, *D*, are defined in Equations 6 and 7, respectively:

$$M_{ij} = (Post_{ij} + Follow_{ij})/2 \tag{6}$$

and

$$D_{ij} = Follow_{ij} - Post_{ij} \tag{7}$$

Although the MANCOVA using *Post* and *Follow* as dependent variables and the MANCOVA using *M* and *D* as dependent variables may appear to be asking different questions, they actually answer the same question "Is there any evidence that the groups differ in any way over time?" and will always provide exactly the same result for the group main effect for a particular data set. Thus, whether the dependent variables used in the MANCOVA are posttest and follow-up or *M* and *D*, the same *F* statistic and *p* value will be obtained when this multivariate test is performed for a given data set.

Table 2 illustrates the conceptual points of this section in the results for our numerical example. Throughout the article, the Type I error rate is set at .05, whereas the confidence interval coverage is set at .95. As expected, the MANCOVA yields the same result, $F(2, 140) = 2.538$, $p = .043$, which is statistically significant, whether *Follow* and *Post* or *D* and *M* are utilized as the dependent variables. Also notice that utilizing the pretest as a linear component of the dependent variables does not change the *F* statistic or *p* value obtained from the MANCOVA when using the pretest as a covariate. Further, none of the MANOVA results yield statistical significance, whereas the MANCOVA does for this hypothetical data set. In the long run, this will typically be the case when comparing MANOVA to MANCOVA because, as mentioned earlier in this section, MANCOVA is generally more powerful than MANOVA when answering the question "Is there evidence that the groups differ in any way over time?" within the context of a randomized PPF design. Even though the MANCOVA approach yields a statistically significant result, allowing us to infer "there is evidence that the groups differ in some way over time," the MANCOVA does not necessarily provide a clear description of where the differences are located. Thus, because the MANCOVA approach does not generally yield a precise determination of where group differences may exist, it is likely that further analyses are needed.

Answering More Specific Questions Within the Context of a Randomized PPF Design

Given the ambiguity of the question answered with MANCOVA, the suggestion here is to usually perform some or all of four different ANCOVAs on the posttest, the follow-up, the *D* variable, and the *M* variable. These four separate analyses have substantively meaningful interpretations as they examine specific types of group differences over time. When any of these four analyses are found to be statistically significant, one is able to conclude that the groups do indeed change dif-

[7]Another plausible analytic method for researchers interested in group differences in any way over time within the context of randomized PPF designs is the omnibus test of the group by time interaction using the repeated measures MANOVA approach. Although this analysis does answer the question "Do the groups differ in any way over time?" it incorporates the pretest as a level of the time factor rather than as a covariate in the model. Because of this, we generally recommend that the MANCOVA be performed when utilizing a multivariate procedure to answer this question within the context of a randomized PPF design.

ferently over time, that is, that there is some group effect on a specific dependent variable.

Although there are four different dependent variables, and thus four different ANCOVAs, that we contend are substantively meaningful, a researcher is not limited to these ANCOVAs nor do all four of these ANCOVAs have to be performed when analyzing data from a randomized PPF design. Rather than simply performing a wide variety of statistical tests, it is best to let theory guide the statistical tests that are performed. By following the suggestions provided here, however, an argument can be made that the recommended approach leads to an increase in the experiment-wise Type I error rate, because four tests are being performed rather than the one test performed in the MANCOVA approach. However, we believe that the four questions can each be thought of as their own distinct family, due to the fact that they all answer qualitatively different questions. Using this recommended approach does not increase the family-wise Type I error rate beyond the nominal α level and does not compromise statistical power by correcting for the change in the experiment-wise Type I error rate. In conclusion, we believe that a researcher will find more substantively meaningful results and less confusion by choosing the ANCOVA(s) that satisfy a researcher's question(s) when analyzing data obtained from a randomized PPF design and using the nominal α level (e.g., an α of .05) for each ANCOVA that is performed.

Group Differences in Change From Pretest to Posttest

Recall that the second question of possible interest is "Do the groups change differently from the pretest to the posttest?" As mentioned in our discussion of the randomized pre–post design, this question is equivalent to "Do the groups differ on the posttest?" when participants are randomly assigned to groups and the pretest is measured prior to the initiation of the treatment. The appropriate analysis for these questions has already been delineated by our discussion of the randomized pre–post design, because the question involves a pretest and one other measurement occasion obtained after the start of the treatment. Thus, ANCOVA on the posttest covarying the pretest is the most appropriate analytic method for answering these questions, whereas the time by group interaction and the group main effect in a one-within, one-between ANOVA, an ANOVA on the difference score, and an ANOVA on the posttest alone are all generally suboptimal analyses in this situation.

Table 3 presents the results for the omnibus tests performed on the posttest and the difference score, *Post – Pre*, using both ANOVA and ANCOVA. As expected, both the ANCOVA on the posttest and the ANCOVA on *Post – Pre* (both analyses using the pretest as a covariate) yield identical results, $F(2, 71) = 3.286$, $p = .043$.[8] All subsequent analyses will report only the results for the ANCOVA on the dependent variable covarying the pretest, as these results pertain to any ANCOVA that covaries the pretest and utilizes the pretest as a linear component of the dependent variable. Also, notice that the ANCOVA yields statistically significant results allowing us to infer that the groups do differ in their change from the pretest to the posttest and, equivalently, the groups are different on the posttest. However, neither ANOVA yields statistical significance in this situation, whereas the ANOVA on the posttest does come close to obtaining statistical significance, $F(2, 72) = 3.090$, $p = .052$. ANCOVA will typically yield a significant result more often than ANOVA when testing for group differences within the context of a randomized design because, as mentioned in the discussion of the randomized pre–post design and shown in Appendixes A and B, ANCOVA is generally a more statistically powerful analytic method than ANOVA within the context of randomized studies.

Although the omnibus tests for both the ANOVA and ANCOVA are illustrated, what is typically of most interest is examining pairwise mean differences between groups (or some other more specific comparisons among groups). In fact, if a researcher is interested in pairwise comparisons, these tests should generally be performed regardless of the results of the omnibus significance test for group differences on the population means, although an appropriate multiple comparison procedure should also be utilized. Using the illustrative example, Table 3 illustrates the confidence intervals[9] for the ANOVA and ANCOVA pairwise comparisons corresponding to group differ-

[8] It is generally not the case that the *p* values for MANCOVA and any of the four ANCOVAs are equal to one another. In this sense, the fact that a *p* value of .043 was obtained for both the MANCOVA and the ANCOVA on the posttest is purely a coincidence.

[9] Recall that each of the four ANCOVAs has been conceptualized as its own family. To control the family-wise Type I error rate, the confidence intervals that are reported throughout the article were calculated by using the Tukey honestly significant difference (HSD) criterion. The Bryant–Paulson procedure is generally more appropriate for pairwise comparisons performed within ANCOVA because the covariate is typically a random variable in practice (see chapter 5 of Maxwell & Delaney [1990] for the details of Tukey's HSD method and Bryant & Paulson [1976] for a discussion of the Bryant–Paulson procedure). Still, for large denominator degrees of freedom and one covariate, the difference between the confidence intervals for the Tukey HSD and the Bryant–Paulson procedure are very small, reflecting no practical difference, as was the case for our illustrative example. Also, as noted by Levin, Serlin, and Seaman (1994), Fisher's least significant difference is generally a more powerful approach than the Tukey HSD when only three groups are of interest. However, because Fisher's least significant difference does not control the family-wise Type I error rate for situations with more than three groups, the Tukey HSD method was chosen as the illustrated method, because it (or some modification of it) generally does control the family-wise Type I error rate for any number of groups when analyzing a complete set of pairwise comparisons.

Table 3. *ANOVA and ANCOVA Omnibus Tests and Pairwise Comparison Confidence Intervals for Posttest (Post) and Posttest Minus Pretest (Post – Pre)*

	Post						*Post – Pre*					
	ANCOVA Covarying Pretest			ANOVA			ANCOVA Covarying Pretest			ANOVA		
Effect	**Lower**	**Upper**	**Width**	**Lower**	**Upper**	**Width**	**Lower**	**Upper**	**Width**	**Lower**	**Upper**	**Width**
Treatment A to Control	–3.024	1.370	4.394	–3.155	1.747	4.902	–3.024	1.370	4.394	–3.643	1.659	5.301
Treatment B to Control	–4.519	–0.125	4.395	–4.922	–0.020	4.902	–4.519	–0.125	4.395	–4.772	0.529	5.301
Treatment A to Treatment B	–0.706	3.696	4.402	–0.684	4.218	4.902	–0.706	3.696	4.402	–1.521	3.780	5.301
Omnibus Test	$F(2, 71) = 3.286$, $p = .043$			$F(2, 72) = 3.090$, $p = .052$			$F(2, 71) = 3.286$, $p = .043$			$F(2, 72) = 1.837$, $p = .167$		

Note: ANOVA = analysis of variance; ANCOVA = analysis of covariance.

ences in change from the pretest to the posttest. Notice that, in each case, the width of the ANCOVA confidence interval is smaller than the corresponding ANOVA interval representing a more precise estimate of the mean difference between the groups. Thus, as we have asserted in this article, the ANCOVA approach is generally the more precise of the two methods.

Focusing on the specific results for ANCOVA from Table 3, there is a statistically significant difference between the Control group and Treatment B because zero (the value corresponding to the null hypothesis of no group mean differences) is not contained in this confidence interval. Because the confidence intervals comparing the Control group to Treatment A and Treatment A to Treatment B both contain zero, neither of these differences is statistically significant. Thus, it has been shown that the Control group and Treatment B differ on their population mean posttest scores and their mean change from the pretest to the posttest, demonstrating Treatment B significantly lowered depression scores when compared to the Control group. It is plausible, however, that the population mean differences between the Control group and Treatment A, as well as the population mean differences between Treatment A and Treatment B, are zero.

Group Differences in Change From Pretest to Follow-Up

The third question that can potentially be answered through a randomized PPF design is "Do the groups change differently from the pretest to the follow-up?" This question is equivalent to "Do the groups differ on the follow-up?" when random assignment to groups is employed and the pretest is measured prior to the initiation of the treatment. The method used to answer this question is identical to the method used to answer the main question of interest in a randomized pre–post design. From a statistical perspective, it does not matter whether the dependent variable is labeled as a posttest or as a follow-up. Thus, the most powerful analysis once again uses the pretest as a covariate in the model and determines whether the groups are significantly different on the follow-up controlling for the pretest. If they are significantly different, one can infer that the groups are different on the follow-up or, equivalently, that the groups do differ in their change from the pretest to the follow-up.

The results for the ANOVA and ANCOVA omnibus tests for group differences on the follow-up are illustrated in Table 4. ANCOVA on the follow-up covarying the pretest yields a statistically significant result, $F(2, 71) = 3.581$, $p = .033$, as does ANOVA on the follow-up, $F(2, 72) = 3.544$, $p = .034$. Again, the ANOVA on the difference score, *Follow – Pre*, fails to reach statistical significance. Thus, in this situation, both ANOVA and ANCOVA on the follow-up reach statistical significance, whereas the ANCOVA has a slightly smaller p value.

The corresponding confidence intervals for the ANCOVA and ANOVA perspectives regarding change from the pretest to the follow-up are also given in Table 4. Notice that for both the ANCOVA and ANOVA on the posttest, it can be inferred that Treatment B has a lower mean than the Control group, whereas all other confidence intervals contain zero, illustrating the corresponding mean differences are not statistically significant. Again notice that all of the ANCOVA confidence intervals are more precise (i.e., more narrow) than the ANOVA confidence intervals for either *Follow* or *Follow – Post*. Focusing on the ANCOVA approach, we can infer that Treatment B lowers depression scores below the Control group's depression scores at the follow-up or, equivalently, Treatment B and the Control group change differently from the pretest to the follow-up (where Treatment B exhibits a greater mean decrease).

Group Differences in Change From Posttest to Follow-Up

The fourth question of interest is "Do the groups change differently from the posttest to the follow-up?" The purpose of this question is ideally to identify group effects during the time period from the posttest to follow-up. Researchers generally are interested in treatment effects for groups that are equivalent in order to

Table 4. *ANOVA and ANCOVA Omnibus Tests and Pairwise Comparison Confidence Intervals for Questions Regarding Change From the Pretest to the Follow-Up, Change From the Posttest to the Follow-Up, and Group Differences on the Average of the Posttest and the Follow-Up*

	Follow						***Follow – Pre***		
	ANCOVA Covarying Pretest			**ANOVA**			**ANOVA**		
Effect	**Lower**	**Upper**	**Width**	**Lower**	**Upper**	**Width**	**Lower**	**Upper**	**Width**
Treatment A to Control	–3.270	1.867	5.137	–3.349	2.180	5.529	–3.860	2.115	5.975
Treatment B to Control	–5.334	–0.195	5.138	–5.671	–0.143	5.529	–5.545	0.431	5.975
Treatment A to Treatment B	–0.510	4.636	5.147	–0.442	5.087	5.529	–1.303	4.672	5.975
Omnibus Test	$F(2, 71) = 3.581$, $p = .033$			$F(2, 72) = 3.544$, $p = .034$			$F(2, 72) = 2.168$, $p = .122$		
	D						***D – Pre***		
Treatment A to Control	–2.726	2.977	5.702	–2.710	2.950	5.659	–4.065	3.729	7.794
Treatment B to Control	–3.294	2.409	5.704	–3.265	2.394	5.659	–3.983	3.811	7.794
Treatment A to Treatment B	–2.289	3.424	5.713	–2.274	3.385	5.659	–3.979	3.815	7.794
Omnibus Test	$F(2, 71) = 0.125$, $p = .883$			$F(2, 72) = 0.122$, $p = .885$			$F(2, 72) = 0.005$, $p = .995$		
	M						***M – Pre***		
Treatment A to Control	–2.682	1.154	3.836	–2.84	1.552	4.392	–3.376	1.512	4.888
Treatment B to Control	–4.462	–0.625	3.837	–4.885	–0.493	4.392	–4.783	0.105	4.888
Treatment A to Treatment B	–0.143	3.701	3.843	–0.151	4.241	4.392	–1.037	3.851	4.888
Omnibus Test	$F(2, 71) = 5.296$, $p = .007$			$F(2, 72) = 4.682$, $p = .012$			$F(2, 72) = 2.659$, $p = .077$		

Note: ANOVA = analysis of variance; ANCOVA = analysis of covariance.

make causal inferences about the differences between the groups after some treatment has been implemented. However, a typical goal in intervention research is to produce groups that are different at posttest due to the treatment that was administered. Therefore, examining group differences in change from posttest to follow-up is a qualitatively different question than examining differences between groups at either the posttest or the follow-up individually. The reason this question is qualitatively different is because the design now potentially compares *nonequivalent* groups, which creates a longstanding methodological conundrum often referred to as Lord's paradox (Lord, 1967). Because groups are likely to differ at posttest, comparing groups from posttest to follow-up is fraught with complications even though the groups are initially equivalent at pretest. Shadish, Cook, and Campbell (2002) provided a thorough discussion of these complications.

In particular, because the groups may differ at posttest, there are two different methods that we explore to answer this question. The two methods we discuss in the following two subsections answer the question regarding group differences from posttest to follow-up in different manners. Because the two models we present may lead to very different conclusions regarding change from posttest to follow-up, it is important for researchers to pay close attention to the recommendations of each of the two methods when examining group change from posttest to follow-up.

Group differences in change from posttest to follow-up: Model I. When the question of interest in a PPF design relates to the differences between scores at posttest and follow-up, one alternative is to use D (see Equation 7) as the dependent variable. The section on statistical power comparisons in randomized PPF designs within this article indicates that the most appropriate analytic method for this question utilizes the pretest as a covariate in an ANCOVA, rather than an ANOVA that ignores the pretest or utilizes the pretest as a linear component of the dependent variable. Thus, the full model that this approach follows is given as follows:

$$D_{ij} = \mu_{D_j} + \beta_{D,Pre}\left(Pre_{ij} - \mu_{Pre}\right) + \varepsilon_{ij} \qquad (8)$$

where μ_{D_j} is the population mean on D for group j, $\beta_{D,Pre}$ is the population regression slope predicting D from the pretest, and ε_{ij} is the error for individual i in group j. This model must be interpreted with the understanding that it is likely that there are differences between the groups at posttest on some concomitant variable(s), the outcome variable being measured, or both. Differences between the groups at the posttest on the outcome variable are taken into consideration by subtracting the posttest from the follow-up within the D variable, a method that restricts the value of the regression slope predicting the follow-up from the posttest to be 1.

It is important to note that this model does not answer the question "If the groups were equal on the outcome variable and all other concomitant variables at the posttest, would their D variables differ from one another?" Thus, if the investigator is interested in answering this question, the method provided here is not useful. Rather the question that is answered by Model I is

"Do the groups change differently from the posttest to the follow-up?" Notice this question does not make a statement about attempting to equalize the groups at the posttest.

Another question that is answered when performing an ANCOVA on the *D* variable covarying the pretest is "Are the magnitudes of the treatment effects the same at the posttest and the follow-up?" Finding a statistically significant omnibus test for the ANCOVA on *D* allows the researcher to infer that the magnitudes of the treatment effects are not the same at the posttest and the follow-up. Further, when forming ANCOVA based confidence intervals for pairwise comparisons on the *D* variable, an inference about the plausible values for the change in magnitude of the treatment effect can be made. Also, an inference about the directionality of the change in the treatment effect's magnitude can be made if the confidence interval does not contain zero. If the confidence interval for the pairwise comparison does contain zero, then it is at least plausible that the magnitude of the treatment effect does maintain itself from the posttest to the follow-up, although it is also plausible that a more statistically powerful study could have detected that the magnitude of the treatment effect is different at these two time points.

The results for the numerical example related to the question of group differences in change from the posttest to the follow-up are show in Table 4. None of the analytic methods come close to reaching statistical significance for this hypothetical data set, whereas the ANCOVA on *D* covarying the pretest provides the lowest *p* value, $F(2, 71) = 0.125$, $p = .883$. The results for the ANOVA on *D* and the ANOVA on *D – Pre* are, respectively, $F(2, 72) = 0.122$, $p = .885$, and $F(2, 72) = 0.005$, $p = .995$. This analysis was unable to show that nonparallelism exists between the groups from posttest to follow-up, and the null hypothesis that groups change equally from posttest to follow-up cannot be rejected.

The corresponding confidence intervals for the ANCOVA and ANOVA on *D* and *D – Pre* perspectives are given in Table 4. Each of the six confidence intervals contain zero, illustrating there is not enough evidence to determine if any of the population mean differences for the groups are statistically significant. Interestingly, the confidence intervals corresponding to the ANOVA on *D* are more precise than the ANCOVA confidence intervals. These results illustrate that the ANCOVA results are not always more precise in the sample, but rather the ANCOVA confidence intervals are generally more precise in the long run.[10]

Group differences in change from posttest to follow-up: Model II. Another alternative for answering the question about group differences from posttest to follow-up utilizes the posttest and the pretest as covariates in the model. The following is the full model for this approach:

$$Follow_{ij} = \mu_{Follow_j} + \beta_{Follow,Pre}\left(Pre_{ij} - \mu_{Pre}\right) + \beta_{Follow,Post}\left(Post_{ij} - \mu_{Post}\right) + \varepsilon_{ij} \quad (9)$$

where μ_{Follow_j} is the population mean score on the follow-up for group *j*, $\beta_{Follow,\ Pre}$ and $\beta_{Follow,\ Post}$ are the population partial regression slopes for the pretest and the posttest respectively, ε_{ij} is the error for individual *i* in group *j,* and μ_{Post} is the population grand mean on the posttest. The specific question that is associated with this model is "Would the mean group change from posttest to follow-up be different had groups been equal on the outcome variable at the posttest?" Notice that this is the only method we discuss in which two time points are covariates in the statistical model. Theoretically, a potential advantage of this model over Model I is that the regression slope predicting the follow-up from the posttest is estimated from the data rather than being restricted to the value of 1. Nevertheless, this analysis fails to provide an answer for the "true" treatment effect, that is, the difference between the groups on the follow-up if the groups had been equal on the outcome variable and all other concomitant variables at the posttest.

A complication arises when Model II is utilized in practice. Because typically the outcome variable at the posttest will be measured with some degree of measurement error, utilizing Model II will generally yield biased estimates of the treatment effect corresponding to Model II (see Huitema, 1980, pp. 111–115, case 3, for the details of this problem). The amount and direction of the bias in the treatment effect obtained from Model II will depend on the amount of measurement error that is present in the posttest measure, typically yielding an estimate of the treatment effect that cannot be trusted to reflect the treatment effect that should be obtained from Model II in practice. Although Model II may be a plausible option in situations in which the outcome variable is

[10] In any given sample, the ANOVA can be statistically significant even though the ANCOVA is not, or the ANOVA-based confidence interval can be narrower than the ANCOVA-based confidence interval, but in the long run the ANCOVA will virtually always be more powerful and precise. Both the ANCOVA and ANOVA approaches are reported in this article for pedagogical reasons. It is not recommended that researchers perform both approaches in practice to see which yields more favorable results (e.g., smaller *p* value, narrower confidence interval). If both approaches are performed in practice and the researcher takes advantage of the more favorable result, the Type I error rate is inflated through performing multiple statistical tests and the empirical confidence interval coverage will be smaller than the nominal level for the set of statistical tests. Researchers should decide a priori which method to use and report the obtained *F* and *p* values as well as confidence intervals from this chosen method only.

measured without error (or measured with a relatively small amount of error), it is not as useful within psychology due to the potentially misleading results that can be obtained because measurement error is typically present in the outcome variable. Thus, we generally do not recommend Model II when answering questions regarding group differences in change from the posttest to the follow-up within the context of a randomized PPF design and do not report results corresponding with the numerical example for this analytic technique.

Group Differences on the Average of Posttest and Follow-Up

The fifth and final question that may be of interest to researchers within the context of a randomized PPF design is "Are the groups different on the average of the posttest and the follow-up?" This question compares the *M* variables (see Equation 6) of the different groups to infer whether group differences are present on the average of the population mean posttest and follow-up scores. From a practical standpoint, researchers might be interested in whether this average score over the final two time points differs as a function of group membership in the population because the test of the average score can sometimes be more powerful than the test of either posttest or follow-up alone. This is more likely to occur when the population group mean trajectories from the posttest to the follow-up are relatively close to being parallel to one another (i.e., the groups' population *D* variables are similar), all other things being equal.

For the hypothetical data set, ANCOVA and ANOVA demonstrate that indeed there were group differences on the *M* variable in Table 4. The ANCOVA yields an $F(2, 71) = 5.296$, $p = .007$, whereas the ANOVA on *M* yields an $F(2, 72) = 4.682$, $p = .012$. The ANOVA on the difference score, $M - Pre$, fails to reach statistical significance. We can conclude that there is a difference between groups on the average of the posttest and follow-up scores in the population with ANCOVA and ANOVA on the *M* variable in this situation, although the ANCOVA yields a smaller *p* value.

The corresponding confidence intervals for the ANCOVA and ANOVA perspectives for the fifth question are contained in Table 4. The confidence interval around the difference between the population means of the Control group and Treatment B does not contain zero when approached from the ANCOVA and ANOVA on *M* perspective. Thus, we conclude that there is a difference between the average of the posttest and follow-up scores between the Control group and Treatment B, with the Control group having a larger mean than Treatment B. Again, as was the case in the results of previous questions, whereas both ANOVA and ANCOVA on *M* yield statistical significance for the difference between the Control group and Treatment B, the ANCOVA yields a more precise confidence interval. In general, this will be the case because ANCOVA is more likely to yield statistical significance when a treatment effect truly exists and narrower confidence intervals when attempting to answer questions about group comparisons within a randomized PPF design.

Results That Appear Contradictory When Analyzing Data From a Randomized PPF Design

In some situations, the results obtained from the analytic methods proposed in this article may appear to be contradictory. For example, suppose a researcher performed a two-group study and also decided to perform the ANCOVAs on the posttest, the follow-up, and the *M* variable. Further suppose a mean difference of 5 was obtained at the posttest corresponding to a *p* value of .10, a mean difference of 5 was obtained at the follow-up corresponding to a *p* value of .12, and a mean difference of 5 was found on the *M* variable corresponding to a *p* value of .04. The researcher might be confused by the fact that the statistical results imply that there is a statistically significant difference on the *M* variable (the average of the posttest and the follow-up) and not on the posttest or the follow-up alone, yet the observed mean differences for all these approaches are equal to one another in this example.

In fact, there is no reason to consider these results to be contradictory, primarily because some statistical tests will be more powerful than other statistical tests depending on the configurations of the population parameters associated with the pretest, the posttest, and the follow-up. Thus, the researcher in this example may be in a situation in which the population group mean trajectories are parallel to one another (thus, the groups are equal on the *D* variable in the population), and when this is the case, it is likely that the ANCOVA on the *M* variable is more statistically powerful than either the ANCOVA on the posttest or the follow-up alone. This same principle applies to other situations, including the comparison of the MANCOVA to the four ANCOVAs when attempting to determine if the groups differ in any way over time. There may be situations in which the MANCOVA yields statistical significance whereas the four-ANCOVA approach does not and vice versa, because the population parameters associated with the measured time points may yield different levels of statistical power for these two procedures. Thus, it is important to remember that these "apparent" contradictions when analyzing data obtained from a randomized PPF design generally represent the fact that the statistical tests being used have different levels of statistical power and are actually not contradictory at all.

Time Effects Within Condition

Conspicuously absent from our presentation is any mention of how to assess changes over time within an individual treatment condition. We have chosen to concentrate on effects that compare groups to one another because causal inferences can be made about the majority of these effects due to random assignment. However, even with random assignment, effects within a group may be difficult to interpret, because these effects are necessarily assessed from the perspective of a single-group design. Campbell and Stanley (1963) described numerous threats to internal validity in such a single-group design. Nevertheless, we acknowledge that understanding effects *within* a group can sometimes provide a valuable context for interpreting differences *between* groups. In such cases, the PPF design reduces to a single-factor within-subjects design with either two levels or three levels of the time factor. In particular, comparisons of scores at two specific time points reduce to a design with two levels of the time factor whereas questions involving all three time points require three levels of the time factor. Standard within-subjects analyses are appropriate to address questions of mean differences over time within a group, although researchers must be sensitive to the likely violation of the sphericity assumption required by the standard mixed-model ANOVA approach for three or more levels. Due to the likely violation of the sphericity assumption, the MANOVA approach to repeated measures (see, e.g., chapter 13 of Maxwell & Delaney, 1990) or HLM is recommended (see, e.g., Raudenbush & Bryk, 2002).

Missing Data

An unfortunate complication in much clinical intervention research is the problem of missing data. The methods we have presented do not directly address problems of missing data. Nevertheless, with random assignment these methods will produce unbiased estimates of the treatment effect as long as the treatments themselves do not affect the presence or absence of obtaining data at pretest, posttest, or follow-up. Unfortunately, in some situations, this may be a strong assumption. Not only does covarying the pretest typically increase power and precision, it also offers increased protection against biased estimates of treatment effects under certain missing data mechanisms, where missingness depends on the pretest score itself. However, in more complicated situations, covarying the pretest does not guarantee obtaining unbiased estimates of the treatment effect. Readers interested in learning more about various methods for estimating treatment effects with missing data are referred to Delucchi and Bostrom (1999), Schafer and Graham (2002), and Sinharay and Russell (2001).

Using HLM to Assess Group Differences Within the Context of a Randomized PPF Design

Our presentation has primarily focused on comparing ANOVA methods to ANCOVA methods with respect to statistical power and precision in the context of a randomized PPF design. However, another plausible option for analyzing data from a randomized longitudinal design is HLM (also known as random coefficients modeling, multilevel modeling, and mixed-effects modeling; Raudenbush & Bryk, 2002). We now briefly explain how this method compares to the methods that have been presented in this article.

An advantage of HLM when analyzing longitudinal data is that this method allows for the specification of individual growth curves over time, whereas traditional ANOVA and ANCOVA methods do not explicitly allow for this specification. Some methodologists argue that understanding group change over time requires understanding individual change over time (Bryk & Raudenbush, 1987; Rogosa, Brandt, & Zimowski, 1982). From this standpoint, the specification of individual growth curves can theoretically lead to a more precise understanding of change over time and also increased power and precision for the test of group differences in certain situations. However, one practical problem that occurs when applying HLM to a randomized PPF design is that the PPF design implies that only three waves of data are collected. Given this restriction, HLM from an individual growth curve perspective is limited in that this method only allows for a straight-line growth model over time.

If the individuals truly do follow a straight-line growth model in the population, there can be a gain in statistical power and precision when comparing HLM to traditional ANOVA and ANCOVA methods.[11] This possible gain in statistical power, however, will be negligible in many practical situations even if the assumption of straight-line growth is met when employing a randomized PPF design. Unfortunately, it is not likely that this assumption will be even approximately true in many situations in the context of a randomized PPF design. This is because the treatment is typically not implemented

[11]This statement assumes that either the measurement occasions are unequally spaced or the HLM analysis utilizes the latent pretest (assuming straight-line growth) as a covariate in the level-2 equation where the slope is the dependent variable (Rausch & Maxwell, 2003; this analysis can currently be carried out in the computer program, HLM5). If neither of these conditions is true, then HLM offers no power advantage over ANCOVA as long as the observed data are balanced (Raudenbush & Bryk, 2002, p. 188; i.e., all participants are measured at each and every time point).

from the posttest to the follow-up for any of the groups within the study. When this is the case, it is likely that the effect of the treatment will diminish to some degree during this time interval resulting in some type of curvilinear trend over time within the treatment groups. In these situations, the HLM analysis utilizing a straight-line growth model generally will yield biased estimates of the treatment effects.[12] Such biased treatment effect estimates can be misleading when attempting to assess group differences in change within the context of a randomized PPF design.

Another option when utilizing HLM within the context of a randomized PPF design is to allow for a saturated fixed effects model and to vary the specifications for the covariances among the time points. This option may be useful for researchers if there is prior knowledge about this covariance matrix that would lead the researcher to believe the model being chosen is correct. However, if the model chosen is incorrect, some bias in the estimate of the treatment effect may be present. Thus, unless the researcher has sufficient confidence that the model being chosen within HLM is correct, the consequences of choosing the wrong model typically outweigh the generally minimal gain in power obtained by specifying the correct model through HLM within the context of a randomized PPF design. This brings us back to our previous recommendation that, unless there are missing data and the missing data mechanism cannot be handled appropriately by standard ANCOVA methods, data obtained from a randomized PPF design should be analyzed using standard ANCOVA methods when interest lies solely in examining group differences.

Summary

We have considered a variety of possible analyses for the randomized PPF design. Arguably our most important point is that the most appropriate analysis or set of analyses depends on the research question(s) of interest. In particular, the research questions should drive the analyses, not vice versa. We have delineated five specific questions likely to be of particular interest in the randomized PPF design and have described the most appropriate analysis for each of these questions. Of these, the three questions that may be of most interest to clinical child and adolescent researchers are (a) "Do the groups differ in change from the pretest to the posttest?" (b) "Do the groups differ in change from the pretest to the follow-up?" and (c) "Do the groups differ in change from the posttest to the follow-up?" ANCOVAs covarying the pretest where the posttest, the follow-up, and the *D* variable are the dependent variables are the analytic methods that should be used when attempting to answer these questions respectively.

The common thread for all the analyses that are utilized to answer questions within the context of a randomized PPF design is that the pretest should be utilized as a covariate in the model, whereas the utilization of the posttest and the follow-up depends on the particular question asked by the researcher. We believe that the three questions just mentioned will generally be the central questions of interest to researchers, whereas the other two questions mentioned in this article will likely cover all remaining questions involving group differences. We also believe that the principles we establish here can be useful to researchers whose specific questions happen to diverge or expand on the specific examples we have demonstrated.

[12]There are some exceptions to this rule. For example, if the time points are equally spaced, the observed data are balanced, and quadratic growth occurs over time for the individuals, then the straight-line growth model will generally provide an unbiased estimate of the treatment effect provided the latent pretest (assuming straight-line growth) is not included as a covariate in the level-2 prediction equation of the slope (Rausch & Maxwell, 2003). However, there is typically no reason to believe that the growth occurring over time is necessarily quadratic rather than some other type of curvilinear trend. In general, the exceptions that provide an unbiased estimate of the treatment effect when utilizing a straight-line growth model for data collected from a randomized PPF design typically rely on assumptions that are not likely to be met in practice.

References

Bryant, J. L., & Paulson, A. S. (1976). An extension of Tukey's method of multiple comparisons to experimental designs with concomitant variables. *Biometrika, 63,* 631–638.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101,* 147–158.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Delucchi, K., & Bostrom, A. (1999). Small sample longitudinal clinical trial with missing data: A comparison of analytic methods. *Psychological Methods, 4,* 158–172.

Friedman, L. M., Furberg, C., & DeMets, D. L. (1998). *Fundamentals of clinical trials* (3rd ed.). New York: Springer-Verlag.

Gatsonis, C., & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin, 106,* 516–524.

Hendrix, L. J., Carter, M. W., & Hintze, J. L. (1979). A comparison of five statistical methods for analyzing pretest–posttest designs. *Journal of Experimental Education, 47,* 96–102.

Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest–posttest design: A potentially confusing task. *Psychological Bulletin, 82,* 511–518.

Huitema, B. E. (1980). *The analysis of covariance and alternatives.* New York: Wiley.

Kenny, D. A. (1979). *Correlation and causality.* New York: Wiley-Interscience.

Levin, J. R., Serlin, R. C., & Seaman, M. A. (1994). A controlled, powerful multiple-comparison strategy for several situations. *Psychological Bulletin, 115,* 153–159.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68,* 304–305.

Matthews, J. N. S. (2000). *An introduction to randomized controlled clinical trials.* London: Arnold.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective.* Belmont, CA: Wadsworth.

Maxwell, S. E., Delaney, H. D., & Dill, C.A. (1984). Another look at ANCOVA versus blocking. *Psychological Bulletin, 95,* 136–147.

Maxwell, S. E., O'Callaghan, M. F., & Delaney, H. D. (1993). Analysis of covariance. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 63–104). New York: Marcel Dekker.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Rausch, J. R., & Maxwell, S. E. (2003). *Longitudinal designs in randomized group comparisons: Optimizing power when the latent individual growth trajectories follow straight-lines.* Manuscript in preparation.

Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin, 88,* 307–321.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92,* 726–748.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton-Mifflin.

Sinharay, S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6,* 317–329.

Appendix A
Effect Size Derivations for ANOVA and ANCOVA

The effect size derivations presented in this appendix assume that participants are randomly assigned to groups and that the pretest is measured prior to the initiation of the treatment. Following Cohen (1988, p. 275; see also Maxwell & Delaney, 1990, p. 101), the population standardized effect size, *f*, for a one-way ANOVA on the dependent variable (*DV*), is defined as

$$f_{DV}^{\text{ANOVA}} = \frac{\sigma_{\text{m}}^{\text{ANOVA on } DV}}{\sigma_{\varepsilon}^{\text{ANOVA on } DV}} \tag{A1}$$

where $\sigma_{\varepsilon}^{\text{ANOVA on } DV}$ is the population standard deviation of the model error (i.e., the square root of the mean square within) for ANOVA on the *DV* assuming homogeneity of variance and $\sigma_{\text{m}}^{\text{ANOVA on } DV}$ is defined as

$$\sigma_{\text{m}}^{ANOVA \text{ on DV}} = \sqrt{\frac{\sum\left(\mu_{DV_j} - \mu_{DV}\right)^2}{a}} \tag{A2}$$

where μ_{DV_j}, $j = 1, 2, \ldots, a$, is the population mean on the *DV* for group j, μ_{DV} is the population grand mean on the *DV*, and a is the number of groups. When ANOVA is performed on a newly defined dependent variable equal to the initial *DV* plus some weight, *A* (positive or negative), multiplied by the pretest (*Pre*),

$$\sigma_{\text{m}}^{ANOVA \text{ on } DV+A*Pre} = \sqrt{\frac{\sum\left(\mu_{DV_j} + A\mu_{Pre_j} - \left(\mu_{DV} + A\mu_{Pre}\right)\right)^2}{a}} \tag{A3}$$

where $\mu_{Pre\,j}$ is the population mean on the pretest for group j. Because participants are randomly assigned to groups and the pretest is measured prior to the initiation of the treatment, the group population means on the *Pre* are equal to one another and thus the population grand mean on the *Pre*:

$$\mu_{Pre_j} = \mu_{Pre} \tag{A4}$$

Equation A4 allows Equation A3 to reduce to Equation A2:

$$\sigma_{\text{m}}^{ANOVA \text{ on } DV+A*Pre} = \sigma_{\text{m}}^{ANOVA \text{ on } DV} \tag{A5}$$

demonstrating that adding a multiple of the pretest to the dependent variable does not affect the numerator of the *f* for ANOVA.

The σ_{m} for ANCOVA (assuming homogeneity of regression slopes) on the *DV* + *A***Pre* is equal to the σ_{m} for ANOVA on the *DV* + *A***Pre* because the adjusted mean on the DV + *A***Pre* is equal to the unadjusted mean on the *DV* + *A**Pre in the population. This can be seen through analyzing the equation for a population adjusted group mean on the *DV* + *A***Pre*, following Maxwell and Delaney (1990, p. 373, equation 27):

$$\mu'_{(DV_j + A*Pre)} = \mu_{DV_j} + A\mu_{Pre_j} - \beta_{DV+A*Pre,Pre}\left(\mu_{Pre_j} - \mu_{Pre}\right) \tag{A6}$$

where $\beta_{DV\,+\,A*Pre,Pre}$ is the population regression slope predicting *DV* + *A***Pre* from *Pre*.

Because of random assignment to groups and the measurement of the pretest before the initiation of the treatment, Equation A4 holds so that Equation A6 reduces to

$$\mu'_{(DV_j + A*Pre)} = \mu_{DV_j} + A\mu_{Pre_j} \tag{A7}$$

The σ_{m} for an ANCOVA on the *DV* + *A**Pre is equivalent to the σ_{m} for an ANOVA on the population adjusted means in Equation A7. Further, the σ_{m} for an ANOVA on the population adjusted means is equivalent to the σ_{m} shown in Equation A3, which was shown to reduce to Equation A2:

$$\sigma_{\text{m}}^{\text{ANOVA on } DV} = \sigma_{\text{m}}^{\text{ANOVA on } DV+A*Pre} = \sigma_{\text{m}}^{\text{ANCOVA on } DV+A*Pre} = \sigma_{\text{m}} \tag{A8}$$

Because the σ_{m} for ANCOVA on the *DV* + *A***Pre* does not depend on the choice of *A*, the equality in Equation

A8 also refers to ANCOVA on the *DV* covarying the pretest when *A* is set equal to zero:

$$\sigma_m^{\text{ANOVA on } DV} = \sigma_m^{\text{ANOVA on } DV+A*Pre} = \sigma_m^{\text{ANCOVA on } DV} = \sigma_m^{\text{ANCOVA on } DV+A*Pre} = \sigma_m \quad \text{(A9)}$$

which demonstrates that all potential ANOVA and ANCOVA analyses that add some multiple (including the situation where *A* is equal to zero) of the pretest to the dependent variable have the same value for σ_m for randomized designs.

The standard deviation of the model error also must be derived for each analytic method in order to form their respective standardized effect sizes. In ANOVA, the model error variance is the variance of the analyzed dependent variable. Thus, σ_ε for ANOVA on the *DV* + *A*Pre* becomes

$$\sigma_\varepsilon = \sqrt{\sigma_{DV}^2 + A^2\sigma_{Pre}^2 + 2A\sigma_{DV,Pre}} \quad \text{(A10)}$$

or the square root of the variance of *DV* + *A*Pre*, where σ_{DV}^2 is the population variance of the initial dependent variable, σ_{Pre}^2 is the population variance of the pretest, and $\sigma_{DV,\,Pre}$ is the population covariance between the initial dependent variable and the pretest. The *f* for ANOVA on the *DV* + *A*Pre* is found by replacing the terms in Equation A1 by Equations A9 and A10,

$$f_{DV+A*Pre}^{\text{ANOVA}} = \frac{\sigma_m}{\sqrt{\sigma_{DV}^2 + A^2\sigma_{Pre}^2 + 2A\sigma_{DV,Pre}}} \quad \text{(A11)}$$

Following the same logic as Equation A10, σ_ε for ANCOVA on the *DV* can be expressed as the square root of the variance of the *DV* unaccounted for by *Pre*:

$$\sigma_\varepsilon^{\text{ANCOVA on } DV} = \sigma_{DV}\sqrt{1-\rho_{DV,Pre}^2} \quad \text{(A12)}$$

where $\rho_{DV,\,Pre}$ is the population correlation between the initial dependent variable and the pretest. Replacing the terms in Equation A1 with Equations A9 and A12, the *f* for ANCOVA on the *DV* can be expressed as

$$f_{DV}^{\text{ANCOVA}} = \frac{\sigma_m}{\sigma_{DV}\sqrt{1-\rho_{DV,Pre}^2}} \quad \text{(A13)}$$

σ_ε for ANCOVA on the *DV* +*A*Pre* can be expressed as the square root of the population variance of the *DV* + *A*Pre* unaccounted for by *Pre* (i.e., the square root of the population partial variance of the dependent variable after controlling for the pretest),

$$\sigma_\varepsilon^{\text{ANCOVA on } DV+A*Pre} = \sqrt{\sigma_{DV+A*Pre.Pre}^2} \quad \text{(A14)}$$

A general equation for the population partial variance of some random variable, *B*, controlling for another random variable, *C*, is

$$\sigma_{B.C}^2 = \sigma_B^2 - \frac{\sigma_{B,C}^2}{\sigma_C^2} \quad \text{(A15)}$$

where σ_B^2 is the population variance of *B*, $\sigma_{B,C}^2$ is the square of the population covariance between *B* and *C*, and σ_C^2 is the population variance of *C*. Then, the population partial variance for *DV* + *A*Pre* controlling for *Pre* can be expressed as

$$\sigma_{DV+A*Pre.Pre}^2 = \sigma_{DV}^2 + A^2\sigma_{Pre}^2 + 2A\sigma_{DV,Pre} - \frac{\left(\sigma_{DV,Pre} + A\sigma_{Pre}^2\right)^2}{\sigma_{Pre}^2} \quad \text{(A16)}$$

By canceling like terms within Equation A16,

$$\sigma_{DV+A*Pre.Pre}^2 = \sigma_{DV}^2 - \frac{\sigma_{DV,Pre}^2}{\sigma_{Pre}^2} \quad \text{(A17)}$$

which reduces to

$$\sigma_{DV+A*Pre.Pre}^2 = \sigma_{DV}^2\left(1-\rho_{DV,Pre}^2\right) \quad \text{(A18)}$$

Placing Equation A18 within Equation A14 yields

$$\sigma_\varepsilon^{\text{ANCOVA on } DV+A*Pre} = \sigma_{DV}\sqrt{\left(1-\rho_{DV,Pre}^2\right)} \quad \text{(A19)}$$

Thus, the *f* for ANCOVA on the *DV* + *A*Pre* can be found by replacing the values in Equation A1 with Equations A9 and A19,

$$f_{DV+A*Pre}^{\text{ANCOVA}} = \frac{\sigma_m}{\sigma_{DV}\sqrt{1-\rho_{DV,Pre}^2}} \quad \text{(A20)}$$

which is equivalent to the *f* for ANCOVA on the *DV* found in Equation A13. Thus,

$$f_{DV+A*Pre}^{\text{ANCOVA}} = f_{DV}^{\text{ANCOVA}} \quad \text{(A21)}$$

Equation A21 shows that as long as the pretest is a covariate in the model, the effect size is not altered by adding (or subtracting) some multiple of the pretest to the dependent variable. Notice the dependent variable could be a posttest score, a follow-up score, an average of posttest and follow-up (i.e., the *M* variable, Equation 6), or the difference between the follow-up and the

posttest (i.e., the *D* variable, Equation 7). Also, although not demonstrated within the previous proof, both analyses illustrated in Equation A21 will yield the same results for significances tests and confidence intervals when comparing group means for a particular data set.

Appendix B
Effect Size Comparison for ANOVA on the *DV + A*Pre* and ANCOVA on the *DV*

For a fixed sample size, the ANCOVA on the *DV* will be more powerful than the ANOVA on the *DV + A*Pre* when

$$f_{DV}^{\text{ANCOVA}} > f_{DV+A*Pre}^{\text{ANOVA}} \tag{B1}$$

This inequality for the subsequent comparison of the ANOVA and ANCOVA standardized effect sizes technically relies on the theoretical assumption that the pretest is fixed across theoretical replications of the study. In practice, it is typically the case that the pretest is actually random over theoretical replications. However, Gatsonis and Sampson (1989) stated that the distribution of the power function for the multiple correlation for random predictors can be approximated well by the power function for the multiple correlation for fixed predictors. In the situation we are interested in, there is only one random predictor and, in the context of regression, the regression coefficient(s) of interest correspond to dummy variables for group status, which are fixed predictors. Therefore, it is likely that the power for testing group mean differences in ANCOVA when the covariate is random is even more closely approximated by the ANCOVA analytic expression used to calculate power that assumes the covariate is fixed than the situation referred to by Gatsonis and Sampson.

Another technical issue with the comparison made in Equation B1 is the loss of one denominator degree of freedom for the ANCOVA effect size due to the estimation of the population regression coefficient predicting the dependent variable from the pretest. However, unless the ANCOVA denominator degrees of freedom are fairly small, this 1 *df* difference is negligible, and the effect size comparison made here is again an accurate approximation.

For example, Maxwell, Delaney, and Dill (1984) noted that the power for the one random covariate case is generally different than the power for the one fixed covariate case. Yet the empirical power values reported in Table 3 of Maxwell et al. (1984, p. 142) for the one random covariate case can be shown to be approximated well by the fixed covariate method of finding the power of ANCOVA that also disregards the loss of 1 *df* in the denominator of the ANCOVA.[13] Thus, although the subsequent derived comparisons are technically approximations that disregard the loss of one denominator degree of freedom in ANCOVA and are derived under the assumption of a fixed covariate, they are accurate approximations of the power comparisons that would likely occur in practice.

Substituting Equations A11 and A13 into Equation B1 yields

$$\frac{\sigma_m}{\sigma_{DV}\sqrt{1-\rho_{DV,Pre}^2}} > \frac{\sigma_m}{\sqrt{\sigma_{DV}^2 + A^2\sigma_{Pre}^2 + 2A\sigma_{DV,Pre}}} \tag{B2}$$

Squaring both sides of Equation B2 yields

$$\frac{\sigma_m^2}{\sigma_{DV}^2\left(1-\rho_{DV,Pre}^2\right)} > \frac{\sigma_m^2}{\sigma_{DV}^2 + A^2\sigma_{Pre}^2 + 2A\sigma_{DV,Pre}} \tag{B3}$$

Multiplying both sides of Equation B3 by

$$\sigma_{DV}^2\left(1-\rho_{DV,Pre}^2\right)\left(\sigma_{DV}^2 + A^2\sigma_{Pre}^2 + 2A\sigma_{DV,Pre}\right)$$

dividing both sides of Equation B3 by σ_m^2, and canceling like terms yields

$$\sigma_{DV}^2 + A^2\sigma_{Pre}^2 + 2A\sigma_{DV,Pre} > \sigma_{DV}^2\left(1-\rho_{DV,Pre}^2\right) \tag{B4}$$

Subtracting the right-hand side of Equation B4 from both sides of the equation and further reduction yields

$$\sigma_{DV}^2\rho_{DV,Pre}^2 + A^2\sigma_{Pre}^2 + 2A\sigma_{DV,Pre} > 0 \tag{B5}$$

Realizing that

$$\beta_{DV,Pre} = \frac{\sigma_{DV,Pre}}{\sigma_{Pre}^2} \tag{B6}$$

and

$$\rho_{DV,Pre} = \frac{\sigma_{DV,Pre}}{\sigma_{Pre}\sigma_{DV}} \tag{B7}$$

[13] The empirical power values for ANCOVA reported in Maxwell et al. (1984, p. 142) are .697, .564, and .480 for population correlations between the dependent variable and covariate of .67, .50, and .28, respectively, where two groups each with a sample size of 12 and a $\delta = 0.8$ (Cohen's definition of a large effect size) were used to obtain these simulated values. The respective analytic power values for ANCOVA that assume a fixed covariate and disregard the loss of 1 *df* are approximately .713, .581, and .497. All of these values are within .017 of the empirical power values reported by Maxwell et al., demonstrating the relative accuracy of the approximation that assumes a fixed covariate and disregards the loss of 1 *df* in the denominator of ANCOVA even for relatively small per group sample sizes.

dividing both sides of Equation B5 by σ^2_{Pre} yields

$$\beta^2_{DV,Pre} + A^2 + 2A\beta_{DV,Pre} > 0 \qquad \text{(B8)}$$

Factoring Equation B8 yields

$$\left(\beta_{DV,Pre} + A\right)^2 > 0 \qquad \text{(B9)}$$

Equation B9 illustrates that

$$f^{\text{ANCOVA}}_{DV} \geq f^{\text{ANOVA}}_{DV+A*Pre} \qquad \text{(B10)}$$

resulting in more statistical power for the ANCOVA on *DV* than the ANOVA on *DV* + *A***Pre* unless *A* equals $-\beta_{DV,\ Pre}$, the negative value of the population regression slope predicting the *DV* from *Pre*. In this special case, the respective values of statistical power for ANOVA and ANCOVA are (approximately) equal to one another, all other factors being equal. From a practical perspective, this derivation proves that ANCOVA on some dependent variable covarying the pretest is almost always more powerful than an ANOVA on the difference score (*DV* – *Pre*), an ANOVA on an averaged score (*DV* + *Pre*), or an ANOVA on the dependent variable alone (*DV*) in practical research settings.