Invited Review

Simple ways to make the results of exercise science studies more informative

Scott J. Dankel

Objectives: To demonstrate some alternate ways of presenting and analyzing pretest-posttest control group designs relative to what is commonly done in exercise science. An emphasis is placed on using simple examples and avoiding statistical jargon to enhance readability for exercise scientists.

Design & Methods: To examine some concerns with how within subject figures illustrate data, statistics to interpret when analyzing pretest-posttest control groups designs, how to analyze studies involving three time points or those including a third factor, and values to use when testing assumptions of statistical tests.

Results & Conclusions: To improve interpretation of data, researchers assessing pretest-posttest control group designs should report the change score and variability of the change score as opposed to only reporting pre-test and post-test variabilities. When performing a 2 × 2 (group by time) mixed ANOVA the interaction term is the only statistic that needs to be interpreted and no follow-up tests are necessary. When assessing a third time point, the most informative follow-up tests to a significant 3 × 2 (time by group) ANOVA involves performing all three 2 × 2 (time by group) interactions to keep the within subject nature of the data. When including a third factor (in addition to the time and group variables), researchers may wish to compute change scores to eliminate the factor of time and allow for the change to be directly assessed. When examining the assumptions of normality and homogeneity of variance, it is important that the change scores meet the assumptions as opposed to the pre-test and post-test measures.

(*Journal of Trainology* 2020;9:43-49)

Key words: change scores ■ error bars ■ pretest-posttest control group design ■ repeated measures ANOVA ■ within subject design

INTRODUCTION

The results section of a manuscript often includes inferential statistics, figures, and tables that display the data in a way that is easy for the reader to interpret. In the exercise science literature, it is very common for researchers to use within subject designs where each participant is measured at multiple time points (e.g. pre-test and post-test). One common design is the pretest-posttest control group design in which a control and experimental group each complete a pre-test and a post-test. This repeated measures design necessitates the data to be displayed differently because each data point for a given individual must be made relative to other data points from that same individual. As it relates to the within subject factor of time, it is important to keep in mind that the pre-test value is not a *response* to the intervention,[1] and therefore, it is the *change* from pre-test to post-test that is of primary interest. The pretest-posttest control group design can be examined numerous ways, and this topic has been covered extensively.[2] In the exercise science literature, it is common for researchers to employ a 2 × 2 mixed ANOVA with a within subject factor of time (pre-test and post-test) and between subject factor of group (control and experimental). A rehashing of what is important to interpret from this analysis and an examination of alternate ways of presenting within subject data are provided. Furthermore, alternate ways of analyzing studies that have more than two time points or studies including a third factor are examined. Lastly, a look at what values must meet the assumptions of normality and homogeneity of variance for the pretest-posttest control group design is provided. The purpose of this manuscript is to examine alternate ways of presenting and analyzing data within the exercise science field that may improve the interpretation of results.

Within subject data reporting

Before getting into analyzing pretest-posttest control group designs, it is important to mention there may more informative ways to display within-subject data. It is very common for authors to only report pre-test and post-test means and variability statistics (i.e. standard deviations or standard errors). Reporting the data this way hides one of the most important pieces of information which is the variability of the *change* from pre-test to post-test. While the reader can get an idea of the mean change from these figures, there is no way to obtain the variability of the change from pre-test to post-test.[3] After all, the change score and standard error of the change score is what is most meaningful when determining whether a given individual would observe an effect, and this is evident in that it is the only data necessary to compute a t statistic for a paired t test. Therefore, it would make sense for authors to report the *change* from baseline and the variability

Received July 17, 2020; accepted August 24, 2020

From the Department of Health and Exercise Science, Exercise Physiology Laboratory, Rowan University, Glassboro, NJ, USA (S.J.D.)

Communicated by Takashi Abe, PhD

Correspondence to: Dr. Scott Justin Dankel, Rowan University, 201 Mullica Hill Rd, Glassboro, NJ 08094, USA
Email: dankel47@rowan.edu

***Journal of Trainology* 2020;9:43-49 ©2012 The Active Aging Research Center http://trainology.org/**

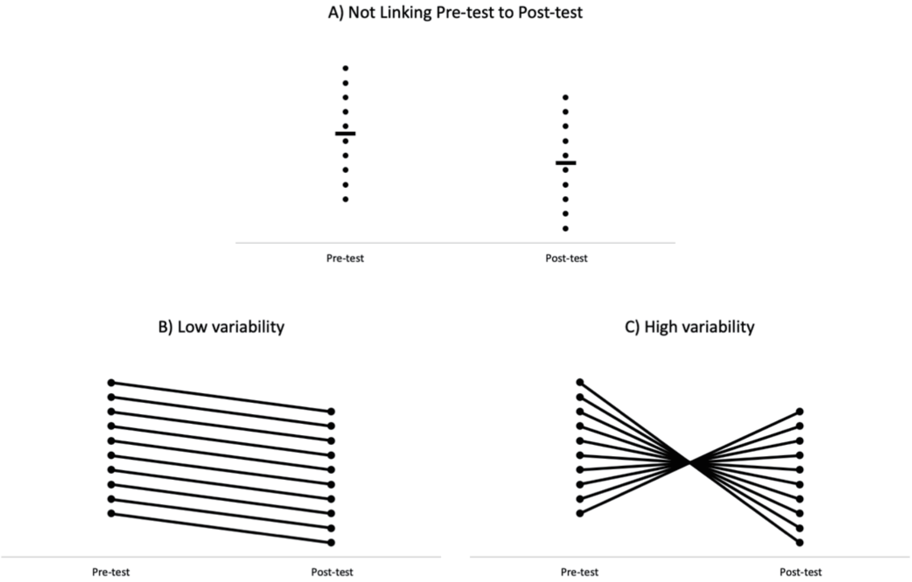


Figure 1 Importance of connecting data points when illustrating paired data. Each of the 3 figures all contain the exact same pre-test and post-test data with the only difference being which data points are paired to one another. A) Showing pre-test and post-test data without connecting data points does not allow the reader to see the variability of the intervention itself. Assuming this were a weight loss intervention: B) illustrates low variability and more confidence that the intervention will work for a given individual. C) illustrates high variability and less confidence that the intervention will work for a given individual. The variability can also be seen if the *changes* from pre-test to post-test are plotted whereby each individual has one data point.

of the *change* as opposed to just reporting the pre-test and post-test variabilities, since the variability of the intervention is often times more important than the variability of the sample itself.[4] Expanding upon this further, it has been recommended that bar graphs not be used for continuous data[5] with the suggestion that authors present individual pre-test to post-test changes particularly when smaller sample sizes are used[6]. It is important that, when using this approach, the authors still present the *change score* across all individuals, or connect pre-test to post-test data points belonging to the same individual. If this is not done, there is no way of determining the variability in response to the intervention (Figure 1). While Figure 1 shows each individual's response to the intervention to provide an idea of the variability and distribution of the change from pre-test to post-test, care should be taken when making inferences on whether these individuals responded differently from one another. Assessing differential responders necessitates a comparison of the change score variability between a control and experimental group, or more appropriately, running multiple interventions.[7]

Analyzing pretest-posttest control group designs

Within the exercise science literature, it is common to examine pretest-posttest control group designs using a 2 × 2 ANOVA which includes a within subject factor of time (pre-test and post-test) and a between subject factor of group (experimental and control). As an example, assume a researcher is trying to test the efficacy of an exercise intervention at reducing body mass. The question of interest is whether the exercise intervention results in a greater reduction in body mass when compared to the control group. The interaction term for the 2 × 2 ANOVA tests just that, which is whether the *change* in body mass differs across groups, and therefore this is the only output that needs to be examined and no follow up tests are necessary.[8] In fact, the p-value for the interaction term on a 2 × 2 ANOVA will be identical to the p-value obtained from an independent t test assessing pre-test to post-test change scores between groups.[9] This all demonstrates the importance of analyzing the interaction term for the 2 × 2 ANOVA or running an independent t test on the change scores to see if the groups *changed differently* from one another. Researchers may also elect to run an ANCOVA on post-test scores while including the pre-test scores as a covariate.[8] For randomized designs, the purpose of the ANCOVA is not necessarily to adjust for baseline differences (since the groups are randomly assigned) but rather to reduce error variance and improve statistical power.[10]

Despite the interaction term directly assessing if the two groups changed differently, it is common for researchers to follow up significant interactions by conducting four additional tests. This includes independent t tests comparing the intervention and control groups at each the pre-test and post-test time points, and paired t test to assess if each the experimental and control groups changed from pre-test to post-test. The approach of running an independent t test at the pre-test time point does not make sense when the groups are randomly assigned because any differences are the result of random chance.[11] Running an independent t test on the post-test scores eliminates the benefit of having a pre-test measure and results in what would be the same result as a post-test only design.

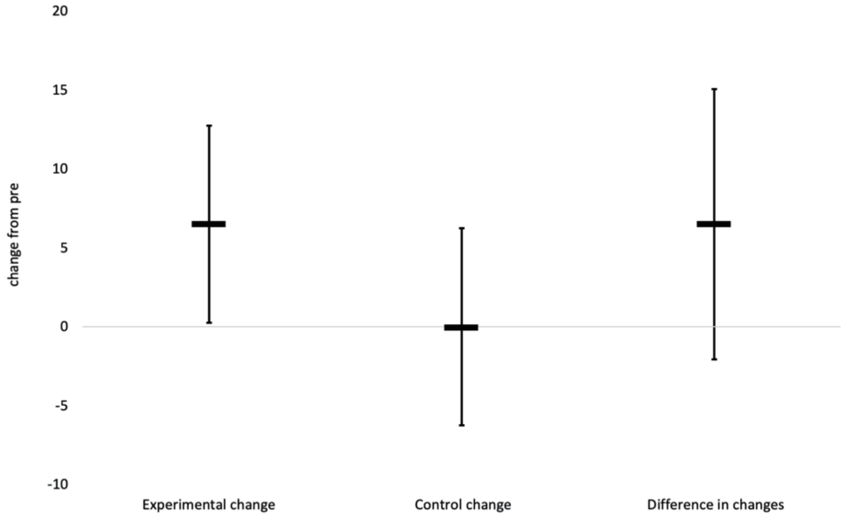


Figure 2 Importance of directly comparing the *change* in the control and experimental groups as opposed to examining separate within group changes. This graph was created assuming 20 individuals in each the control and experimental group with each group having a standard error of 3. The experimental group has a change of 6.5 units and the control group has a change of 0 units. The experimental group changes significantly from pre to post, but the control group does not. Despite the difference in statistical significance, there are no significant differences between the groups when compared directly. The values are expressed as means and 95% confidence intervals.

This will likely result in a loss of statistical power as between as opposed to within subject variability is being compared,[12] and pre-test to post-test correlations are often high in the exercise science literature. In other words, comparing only the post-test scores may be problematic because an individual's post-test score will likely be dependent upon what their pre-test score was. For example, an individual who is severely overweight before an intervention may lose a lot of weight, but they will still weigh more at the post-test measure in comparison to a normal weight individual who did not lose any weight. Thus, this does not appropriately assess the intervention itself. One exception is when the pre-test value is used as a covariate. In this case, raw values at the post-test measure or change scores from pre-test to post-test can be used as they will yield the same results.[13]

The within group assessments (paired t tests from pre to post) are also problematic because the groups are never directly compared.[14] In other words, if the control group does not change from pre-test to post-test, but the experimental group does, this still does not mean that the groups changed *differently* (Figure 2). This points to another common mistake in that often times researchers will power studies for within group effects (i.e. to detect a change in the intervention group), but then run an analysis that compares between group effects (comparing the change in an intervention group to the change in a control group), thus resulting in an underpowered study.[15] This is because the error of the intervention group may exceed zero resulting in within subject significance, but it may not exceed the error of the control group, thus not resulting in between group significance (Figure 2).

When no interactions are present, it is common for researchers to then examine main effects. In the example used previously, this would include examining main effects of time and group. Again, all of the information needed to make a conclusion on the efficacy of the intervention is provided in the interaction term when using the pretest-posttest control group design, and little if any useful information is provided by examining the main effects.[8] Huck and McLean[8] state that the main effect of time is "worthless from an experimental point of view" and that the main effect of group "underestimates the variability of the treatment effects" concluding that the interaction term is all that should be analyzed. This is because the main effect of time examines whether individuals changed from pre-test to post-test independent of which group they were assigned to (i.e. collapsing the groups together). Therefore, this cannot provide any useful information about the intervention if the control and experimental groups are not differentiated from one another. Similarly, the main effect of group examines whether the control and experimental groups differ independent of the time point being examined (i.e. collapsing the time points together). Therefore, the main effect of group dilutes the importance of the intervention by including pre-test scores that have not yet been subjected to a treatment.

In the event that two experimental groups are analyzed in the absence of a control group, the emphasis should again be placed on the interaction term and caution should be taken when interpreting the main effect of time. In the absence of an interaction, a main effect of time may seem to indicate that both interventions were equally effective, but the change from pre-test to post-test may have resulted from unexpected effects impacting both groups. Thus, control groups are always recommended. Occasionally, authors in the exercise science literature have elected to use control groups from previously conducted studies, but this does not control for unknown factors specific to the experiment and is generally not recommended as this approach may produce unreliable results.[16]

Analyzing three time points

Studies may sometimes wish to compare a 3 × 2 mixed measures ANOVA using the same between subject factor of group (experimental and control), but now including a third time point for the within subject factor (i.e. pre-test, mid-test and post-test). The main focus will again be on the interaction term to see if the groups changed *differently* over time. If there is an interaction, this indicates that the experimental and control groups changed differently over time. One of the most common approaches to follow up this significant interaction involves running an independent t-test to compare groups at each time point, and a one-way repeated measures ANOVA across time within each group. As mentioned previously, this approach of running independent t tests at a given time point is problematic as the benefit of having within subject data is lost. That is, the researcher is only comparing groups at a given time point as opposed to comparing the *change* from a previous time point. Furthermore, the within subject changes over time within each group do not provide information on whether the groups changed *differently* from one another.[14,17] As such, a significant 3 × 2 interaction may be present (indicating that the groups changed differently over time) but follow-up tests don't show any differences between groups at any time point. This is possible since the denominator of the test statistic would be utilizing the pooled variability on each sample at a given time point (i.e. between subject variability) as opposed to the pooled variability on the change from baseline within each person (i.e. the within subject variability). A more appropriate approach would be to follow up a significant 3 × 2 interaction by analyzing each of the three possible 2 × 2 interactions. The interaction term for each of these ANOVAs will directly compare whether the *changes* differed between groups[9] and the following 2 × 2 ANOVAs can be computed:

- A within subject factor of time (pre-test and mid-test) by between subject factor of group (intervention and control). A significant interaction here would illustrate there is a difference in the change from pre-test to mid-test across groups.
- A within subject factor of time (pre-test and post-test) by between subject factor of group (intervention and control). A significant interaction here would illustrate there is a difference in the change from pre-test to post-test across groups.
- A within subject factor of time (mid-test and post-test) by between subject factor of group (intervention and control). A significant interaction here would illustrate there is a difference in the change from mid-test to post-test across groups.

Researchers can obtain the same follow-up results by comparing independent t tests across groups on the change scores from pre-test to mid-test, pre-test to post-test, and mid-test to post-test.

Analyzing a third factor

The previous example using a 2 × 2 ANOVA has two different factors each with two levels. A researcher may wish to include a third factor into the analysis. Instead of the previous example which examined a 2 × 2 design including a within subject factor of time (pre-test and post-test) and a between subject factor of group (experimental and control), a researcher may want to include a third factor of sex (male and female). Most researchers in exercise science would likely examine this by performing a 2 × 2 × 2 ANOVA consisting of three factors (time, group, and sex) each of which have two levels. If there is a significant interaction with a 3-way ANOVA this can be difficult to analyze since this indicates there is a two-way interaction which varies across levels of the third variable. Researchers could run follow up tests to then isolate levels of one variable and examine the 2-way interaction at each level of the isolated variable. For example, researchers could isolate the time variable and examine the group by sex interaction at the pre-test time point and the group by sex interaction at the post-test time point. The problem here becomes that the within subject nature of the time variable (i.e. the actual change resulting from the intervention) is lost. In other words, the group x sex interaction at pre-test and the group x sex interaction at post-test is examined, but the group x sex interaction on the change from pre-test to post-test is not examined. Therefore, a simple adjustment would be to eliminate the time variable by computing a change score and then running a 2 × 2 (group by sex) ANOVA on the change score. Assuming a researcher is again examining the efficacy of a weight loss intervention (experimental and control group) for reducing body mass, it will now be assumed that the researcher also wants to know if the efficacy of the weight loss intervention depends upon the sex of the individual (male and female). A 2 × 2 (between subject factors of group and sex) ANOVA ran on the change score (i.e. a created variable as post-test body mass – pre-test body mass) would provide the necessary information as listed in the footnote of Figure 3. This could also be examined with a 2 × 2 ANCOVA assessing post-test values or gain scores whilst including the pre-test value as a covariate and the results would be interpreted similar to what is listed in Figure 3. In this example, the interaction term and the main effect of group would be important to interpret when assessing the efficacy of the intervention. A significant interaction would indicate that the effectiveness of the intervention is dependent on the sex of the individual and a significant main effect of group would indicate that the intervention produced a differential change when compared to the control group, with this effect being independent of the sex of the individual. A main effect of sex would indicate that males had a differential change when compared to females, but this occurred independent of group assignment (experimental vs. control). As such, the main effect of sex may not be of interest here as it does not take into account the intervention.

Considerations when using change scores

When using change scores as the dependent variable it is

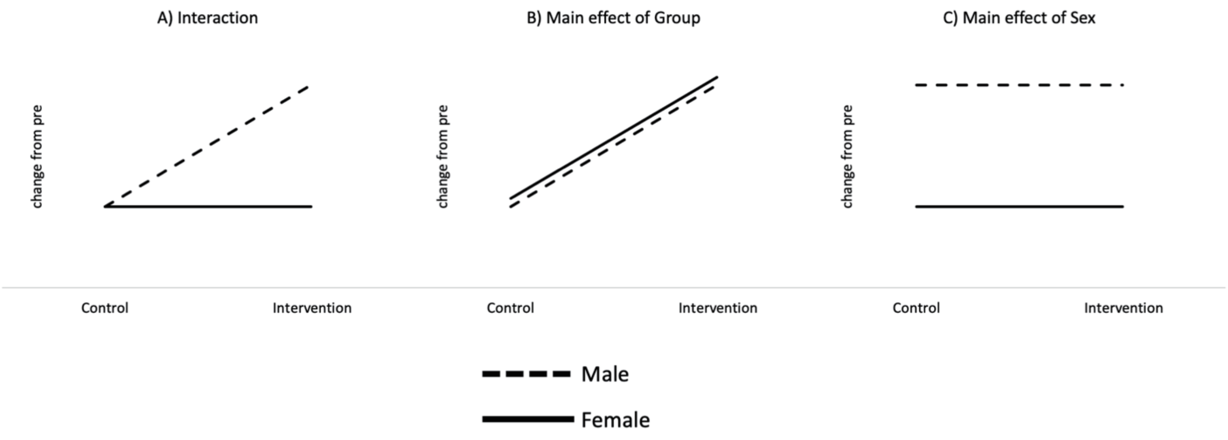
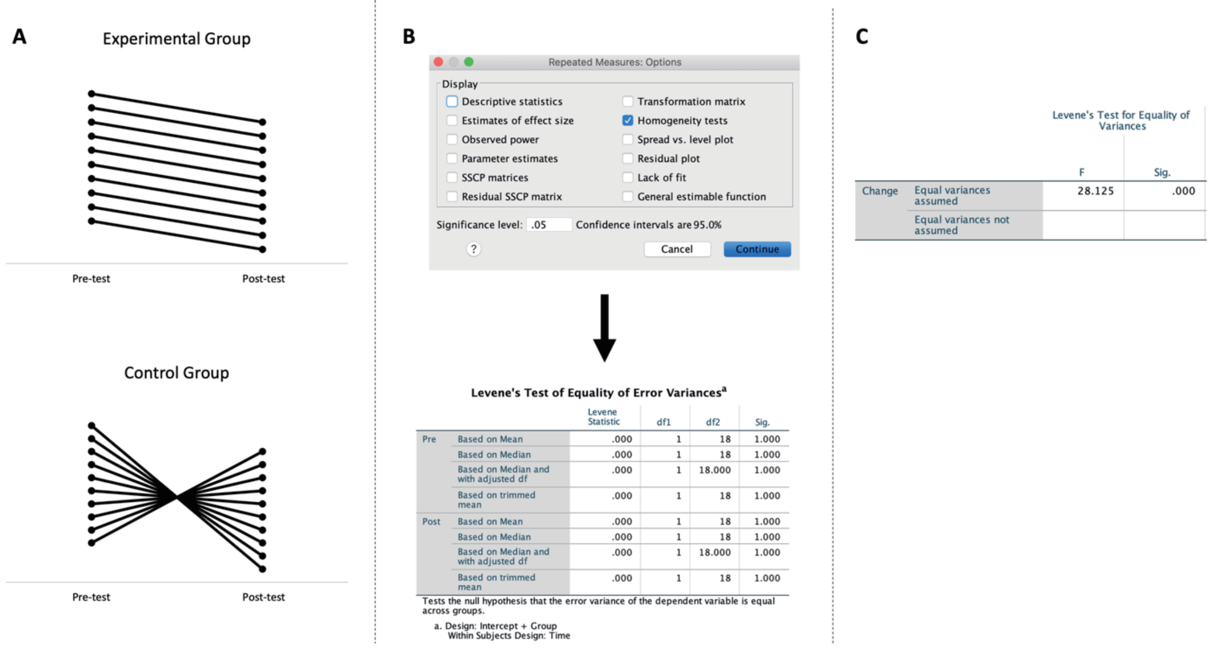


Figure 3 Using a 2×2 ANOVA to compare changes between experimental and control groups when an additional factor is included. In this example, the two factors include the between subject factors of group (intervention and control) and sex (male and female). These figures are not made from the same data set as they are intended to illustrate what may be concluded based on the results of the test. The dependent variable is the change from pre-test to post-test. This can be done to avoid a three-way interaction by using the change score and eliminating time as an additional factor. A) a significant interaction would indicate that the effectiveness of the intervention depended upon the sex of the individual. B) a main effect of group would indicate that the intervention group produced a differential change in body mass when compared to the control group, independent of sex. C) a main effect of sex would indicate that one sex had a differential change in body mass when compared to the other, independent of whether the individual was in the control or experimental group.

important to consider when it may be more appropriate to control for baseline values using an ANCOVA. Change score analyses can be impacted by regression to the mean and therefore do not account for baseline imbalances between groups.[18] Regression to the mean refers to a negative correlation between baseline scores and change scores, such that individuals with lower scores at baseline will likely observe higher change scores. This makes baseline imbalances potentially problematic, and while baseline imbalances are not typically a concern when individuals are randomly assigned to groups, it can be when smaller sample sizes or improper randomization is employed. Alternatively, using an ANCOVA to adjust for baseline values, can negate the influence of regression to the mean and account for baseline imbalances.[19] An ANCOVA will often have more statistical power than using an unadjusted change score analysis, but one must check to make sure the appropriate statistical assumptions are met. In general, an ANCOVA adjusting for baseline values is recommended as it is less likely to result in biased estimates.[1]

Assumptions of statistical tests

When employing the pretest-posttest control group design, researchers often report the results of tests assessing if the statistical assumptions of normality and homogeneity of variance are met. Since researchers are concerned with the change from baseline, it is the *change score* that needs to be normally distributed when examining the efficacy of an intervention.[20] The distribution of the pre-test and post-test scores does not provide information about the distribution of the *change* from pre-test to post-test. This can be seen in Figure 1 where, despite having the same pre-test and post-test distributions, the change scores (assessed via Shapiro-Wilk test) in Figure 1B are not normally distributed ($p < 0.001$), but the change scores in Figure 1C ($p = 0.892$) are. When checking to see if the normality assumption is met, researchers can compute a variable that is the change score and test to see if this variable is normally distributed. Additionally, when comparing a pretest-posttest control group design, the *change scores* must meet the homogeneity of variance assumption, since it is the *change scores* that are being compared with the interaction term.[21] This is demonstrated in Figure 4 using the data presented in Figure 4a. When running a 2×2 (group by time) mixed ANOVA in SPSS (a common software used in the exercise science literature) the option to test for homogeneity of variances is provided (Figure 4b top image). Clicking this box, however, tests if the variance of the pre-test and post-test scores differ across groups (Figure 4b bottom image), but this does not test if the *change score* variance differs across groups (this is tested in Figure 4c). Thus, the pre-test and post-test scores can have identical variances across groups (Figure 4b bottom image), but the change score variance between groups may be quite large (Figure 4c). Since the pre-test-posttest control group design only relies on interpreting the interaction term, the focus of the assumptions should be on the change scores between groups. Therefore, each of the change scores for the control and experimental groups should be normally distributed, and each of these change scores should have approximately equal variances (SPSS will automatically provide homogeneity of variance results when running an independent t-test on change scores). If a third factor is included and a 2×2 (e.g. group x sex) ANOVA on change scores is employed, then each of the 4 groups should be normally distributed and have approximately equal variances (a homogeneity of variance test is provided appropriately in



Repeated Measures: Options

Display

☐ Descriptive statistics
☐ Estimates of effect size
☐ Observed power
☐ Parameter estimates
☐ SSCP matrices
☐ Residual SSCP matrix
☐ Transformation matrix
☑ Homogeneity tests
☐ Spread vs. level plot
☐ Residual plot
☐ Lack of fit
☐ General estimable function

Significance level: .05 Confidence intervals are 95.0%

Levene's Test of Equality of Error Variances[a]

		Levene Statistic	df1	df2	Sig.
Pre	Based on Mean	.000	1	18	1.000
	Based on Median	.000	1	18	1.000
	Based on Median and with adjusted df	.000	1	18.000	1.000
	Based on trimmed mean	.000	1	18	1.000
Post	Based on Mean	.000	1	18	1.000
	Based on Median	.000	1	18	1.000
	Based on Median and with adjusted df	.000	1	18.000	1.000
	Based on trimmed mean	.000	1	18	1.000

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

[a] Design: Intercept + Group
Within Subjects Design: Time

		Levene's Test for Equality of Variances	
		F	Sig.
Change	Equal variances assumed	28.125	.000
	Equal variances not assumed		

Figure 4 Importance of using the change score from pre-test to post-test when testing assumptions of statistical tests for a pretest-posttest control group design. A) Data used to complete the homogeneity tests. B) The top figure illustrates that, when performing a 2×2 ANOVA (within subject factor of time and between subject factor of group), SPSS allows for the option to test for homogeneity of variance. The bottom figure illustrates that this will test the homogeneity of variance on the pre-test and post-test scores which in this example are identical between groups. As the interaction term is what is important, and provides the same p-value as an independent t test on change scores, the appropriate homogeneity of variance test assesses if the variance of the change scores are equal across groups. This information is not provided here. C) SPSS will automatically provide a homogeneity of variance test for independent t tests. Here it indicates that the variance on the change score between groups is not equal, and thus the assumption of homogeneity of variance is violated necessitating a more conservative p-value. Analyses were computed using SPSS version 26.

SPSS using a single f statistic comparing the variability across groups).

CONCLUSION

The pretest-posttest control group design is commonly used in the exercise science literature to test the efficacy of interventions. When these designs are performed, it is important to illustrate the change and the variability of the change as opposed to only reporting the pre-test and post-test variabilities. Doing this allows the reader to interpret the variability of the intervention itself as opposed to the variability of the sample that was recruited. The pretest-posttest control group design can be appropriately tested by performing a 2×2 (time by group) ANOVA and examining the interaction term or by computing an independent t test on the changes from pre-test to post-test as both will yield the same result. If an interaction is found on a 2×2 ANOVA, no follow up tests are necessary. If a third time point is included and researchers have a significant 3×2 (time by group) interaction, it is most appropriate to perform all possible 2×2 (time by group) ANOVAs to see where the groups *changed differently*. If a third factor is included (in addition to the time and group factors), it may be beneficial to run a 2×2 ANOVA on the change scores so the within subject nature of the data is still maintained. Finally, it is important to note that the assumptions of normality and homogeneity of variance for the pretest-posttest control group design are related to the changes from baseline as opposed to the pre-test and post-test values themselves. The intent of this manuscript is to improve the reporting of results within the exercise science literature.

CONFLICT OF INTEREST

The author declares no conflicts of interest.

REFERENCES

1. Senn S. Change from baseline and analysis of covariance revisited. *Stat Med* 2006; 25: 4334-4344.
2. Dugard P, Todman J. Analysis of Pre-test-Post-test Control Group Designs in Educational Research. *Educational Psychology* 1995; 15: 181-198.
3. Weissgerber TL, Milic NM, Winham SJ, et al. Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS Biol* 2015; 13: e1002128.
4. Dankel SJ, Loenneke JP. Effect Sizes for Paired Data Should Use the Change Score Variability Rather Than the Pre-test Variability. *J Strength Cond Res* 2018, Online ahead of print doi: 10.1519/JSC.0000000000002946.
5. Weissgerber TL, Winham SJ, Heinzen EP, et al. Reveal, Don't Conceal: Transforming Data Visualization to Improve Transparency. *Circulation* 2019; 140: 1506-1518.
6. Weissgerber TL, Garovic VD, Savic M, et al. From Static to Interactive: Transforming Data Visualization to Improve Transparency. *PLoS Biol*

2016; 14: e1002484-e.

7. Hecksteden A, Kraushaar J, Scharhag-Rosenberger F, et al. Individual response to exercise training - a statistical perspective. *J Appl Physiol* 2015; 118: 1450-1459.
8. Huck SW, McLean RA. Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychol Bull* 1975; 82: 511-518.
9. Maxwell SE, Howard GS. Change Scores—Necessarily Anathema? *Educ Psychol Meas* 1981; 41: 747-756.
10. Dimitrov DM, Rumrill JPD. Pretest-posttest designs and measurement of change. *Work* 2003; 20: 159-165.
11. de Boer MR, Waterlander WE, Kuijper LDJ, et al. Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate. *Int J Behav Nutr Phys Act* 2015; 12: 4.
12. Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. *BMC Med Res Methodol* 2001; 1: 6.
13. Laird N. Further Comparative Analyses of Pretest-Posttest Research Designs. *The American Statistician* 1983; 37: 329-330.
14. Bland JM, Altman DG. Comparisons within randomised groups can be very misleading. *BMJ* 2011; 342: d561.
15. Brysbaert M. How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. *J Cogn* 2019; 2: 16.
16. Moser P. Out of Control? Managing Baseline Variability in Experimental Studies with Control Groups. In: Bespalov A, Michel MC, Steckler T, editors. Good Research Practice in Non-Clinical Pharmacology and Biomedicine. Cham: *Springer International Publishing*; 2020; pp. 101-117.
17. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. *BMJ* 2003; 326: 219.
18. Vickers AJ, Altman DG. Statistics notes: Analysing controlled trials with baseline and follow up measurements. *BMJ* 2001; 323: 1123-1124.
19. Clifton L, Clifton DA. The correlation between baseline score and post-intervention score, and its implications for statistical analysis. *Trials* 2019; 20: 43.
20. Kim TK. T test as a parametric statistic. *Korean J Anesthesiol* 2015; 68: 540-546.
21. Zientek L, Nimon K, Hammack-Brown B. Analyzing data from a pretest-posttest control group design: The importance of statistical assumptions. *Eur J Training Dev* 2016; 40: 638-659.