2023 ■ Vol. 19 ■ no. 1

Analysis of Treatment-Control Pre-Post-Follow-up Design Data

Donald Sharpe [a] and Robert A. Cribbie [b]

[a] Department of Psychology, University of Regina
[b] Department of Psychology, York University

Abstract ■ The treatment-control pre-post-follow-up (TCPPF) design is a popular means to demonstrate that a treatment group is superior to a control group over time. The TCPPF design can be analyzed using traditional methods (e. g., between-within ANOVA) or with modern multilevel (also known as mixed or hierarchical) modeling. In spite of TCPPF's widespread popularity, there is sparse and confusing guidance for applied researchers on how to analyze data from TCPPF designs using SPSS, one of the most popular software packages for data analysis. We present an introductory tutorial on methods for analyzing TCPPF data. Advantages, disadvantages, and cautions related to applying these approaches are discussed.

Keywords ■ SPSS, ANOVA, Multilevel Models, Hierarchical Models, Mixed Models, Statistics. **Tools** ■ R, SPSS.

✉ sharped@uregina.ca

10.20982/tqmp.19.1.p025

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers ■ Two anonymous reviewers.

Introduction

Randomly assigning participants to treatment and control conditions and then measuring change over time is a popular research strategy in psychology. Pick up any issue of the *Journal of Consulting and Clinical Psychology*, for example, and one will see psychologists are often interested in whether a treatment group differs from a control group over time such as from pre-test to post-test to follow-up. This treatment-control pre-post-follow-up (TCPPF) design has many of the features of the randomized controlled trial and it resembles a basic version of the classic split-plot experiment (see Jones & Nachtsheim, 2009). In its simplest form, one factor of the TCPPF is a between subject factor with two levels, such as a treatment compared to a control group. Ideally, assignment to the treatment and control group is random, although circumstances sometimes make that impossible. Much of what we say here applies regardless of whether or not assignment to the treatment and control group is random. What we have to say also applies to a comparison of two treatments, a control group and multiple treatments, etc. The other factor is a within subject (repeated measures) factor with two or more levels. Of utmost interest is the interaction between the between and within subject factors (Jaccard & Guilamo-Ramos, 2002a); for example, does the treatment group do better than the control group over time?

Although there are many excellent books, journal articles, and websites devoted to advanced statistical analyses, few of these sources focus specifically on TCPPF designs, and those sources that do provide insufficient detail or contradictory advice regarding best practices for data analysis. As stated by Howell (n.d.), "just about every source you read on these models takes a somewhat different approach, and it is not always clear how they relate to each other and why they look at the models so differently" (para 2). Howell also commented that these sources use a variety of statistical software (e. g., SPSS, SAS, R) and he notes that "a discussion written for SAS looks, on the surface, quite different from one written with respect to R" (para 3). SPSS is the statistical package most widely used by researchers in psychology (Davidson et al., 2019), but we searched in vain for a clear step by step guide for how to analyze the TCPPF design using SPSS. Although our focus here is on SPSS, in the supplementary material we include the syntax for doing the same calculations in R.

We received a number of suggestions for resources to analyze TCPPF designs. Peugh and Enders (2005) focus on multilevel analysis in SPSS, but not treatment-control group designs. Their cross-sectional example uses data from 160 public and private schools predicting math achievement from socioeconomic status. Their longitudinal example looks at change over time in a cognitive task, but not as a function of a treatment. Peugh and Ender's discussion of moving from what we will call a marginal model to a true multilevel model is limited to changing the covariance structure and readers are directed by Peugh and Endler to other sources for further explanation. Another suggested source was Heck et al. (2014), who examine multilevel models in SPSS, but experimental designs were only briefly mentioned. Similarly, Field (2018) does an exceptional job of explaining multilevel models in one chapter of his best-selling book, however Field's focus is on what buttons to click in the SPSS MIXED menu, whereas running the analysis via syntax may be more efficient and revealing. Field does devote some attention to the MLM (multilevel modeling) output (when many authors do not), but we feel more interpretation should be provided. Field also performs a random slope analysis that may not be appropriate for many TCPPF designs for a reason we will discuss.

Design Issues

Why Does Randomization Matter? Although our focus is on the statistical analysis of TCPPF data, the research design that produces those data plays a critical role in the choice of statistical analyses. There must be at minimum two groups (e. g., a treatment group and a control group, or two different treatment groups). The nature of the groups is not relevant here (Guidi et al., 2018; Kendall et al., 2013); what is relevant is how participants are assigned to groups. Only through true random assignment can one expect groups to be equivalent at baseline. In their review of 2017 *Journal of Personality and Social Psychology* articles, Chester and Lasko (2021) found 62% of between-participant manipulations were said by the authors of the articles to be the result of random assignment to conditions. Yet none of the authors of those articles reported how participants were randomly assigned to conditions, thus providing no means to assess whether systematic biases might have arisen in that process. Further, in order to have the opportunity to classify factors as *causal*, it is necessary to start with random assignment to conditions. These decisions can have important consequences on how we interpret the findings of the study.

What about Pre-Test Covariates? Although not useful for verifying random assignment success, examination of pre-test scores for non-outcome (i.e., not dependent) variables can help to identify potential covariates. The issue is how that examination should precede. The practice of selecting covariates based on their statistical significance through tests of group differences is frowned upon (Assman et al., 2000; Eghewale, 2015). Rather, covariates selected judicially based on theoretical considerations, covariates strongly correlated to the outcome variable (e. g., moderate effect size), covariates not strongly correlated to each other, and covariates measured before the treatment is introduced, can serve to reduce error variance and increase statistical power (Kahan et al., 2014; Streiner, 2019).

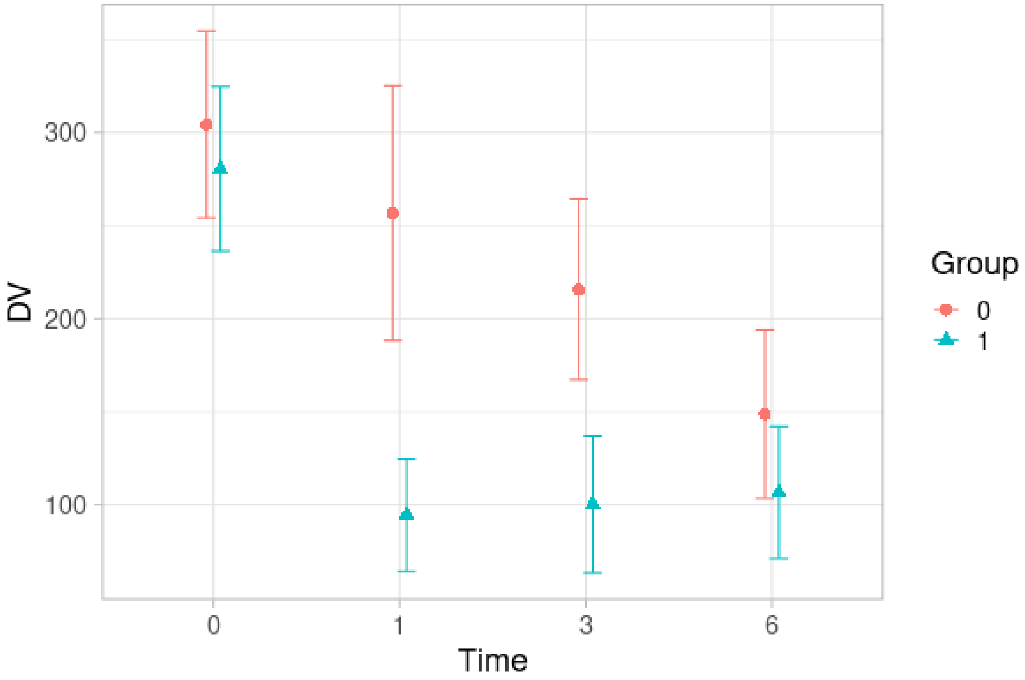
Again, it is worth reiterating the importance of random assignment to groups. When participants are assigned randomly to groups, covariate analysis addresses extraneous and irrelevant variance. Conversely, when participants are *not* assigned randomly assigned to groups, covariate analysis is most tempting to employ but works least well (Streiner, 2019). When group assignment is not random, a covariate will be related to the outcome variable (which is good), but the covariate will also be related to group membership (which is very bad). If related to group membership, a covariate can artificially inflate or deflate differences between groups depending on the nature of the association (Cribbie & Jamieson, 2000). Seeking to equalize groups on a covariate in the absence of random assignment is asking statistics to do the impossible.

How about the Pre-Test Score Serving as a Covariate? This question highlights a longstanding debate (see Jennings & Cribbie, 2016; Rausch et al., 2003) as to whether pre-test to post-test data should be analyzed via gain score analysis (i.e., computing pre-test to post-test difference scores and then examining those difference scores by group through a one-way ANOVA) or via covariance analysis (i.e., comparing the post-test scores between conditions after covarying the pre-test scores via ANCOVA). Only when there are differences between treatment and control groups on the pre-test will gain score and covariance approaches produce meaningfully different results (Counsell & Cribbie, 2017; Cribbie & Jamieson, 2000). Alternatively, one could calculate a mixed (between-within) model ANOVA. A mixed model ANOVA examines differences between groups collapsing over time periods (a between-subject factor), differences between time periods collapsing over groups (a within-subject factor), and the interaction of *Group* by *Time*. For pre-test to post-test data, the interaction term from the mixed model ANOVA is equivalent to the outcome from the gain score approach (Knapp & Schafer, 2009; O'Connell et al., 2017).

The ANOVA Approach

The traditional way to analyze TCPPF data is the mixed model ANOVA. *Mixed* signifies that *Group* is a between (independent) groups variable and *Time* is a within groups (repeated measures) variable. One question answered through a mixed model ANOVA is do groups differ in their change from pre-test to post-test. However, some have suggested that the mixed model ANOVA be abandoned given

Figure 1 ■ Means Plot with 95% Confidence Intervals



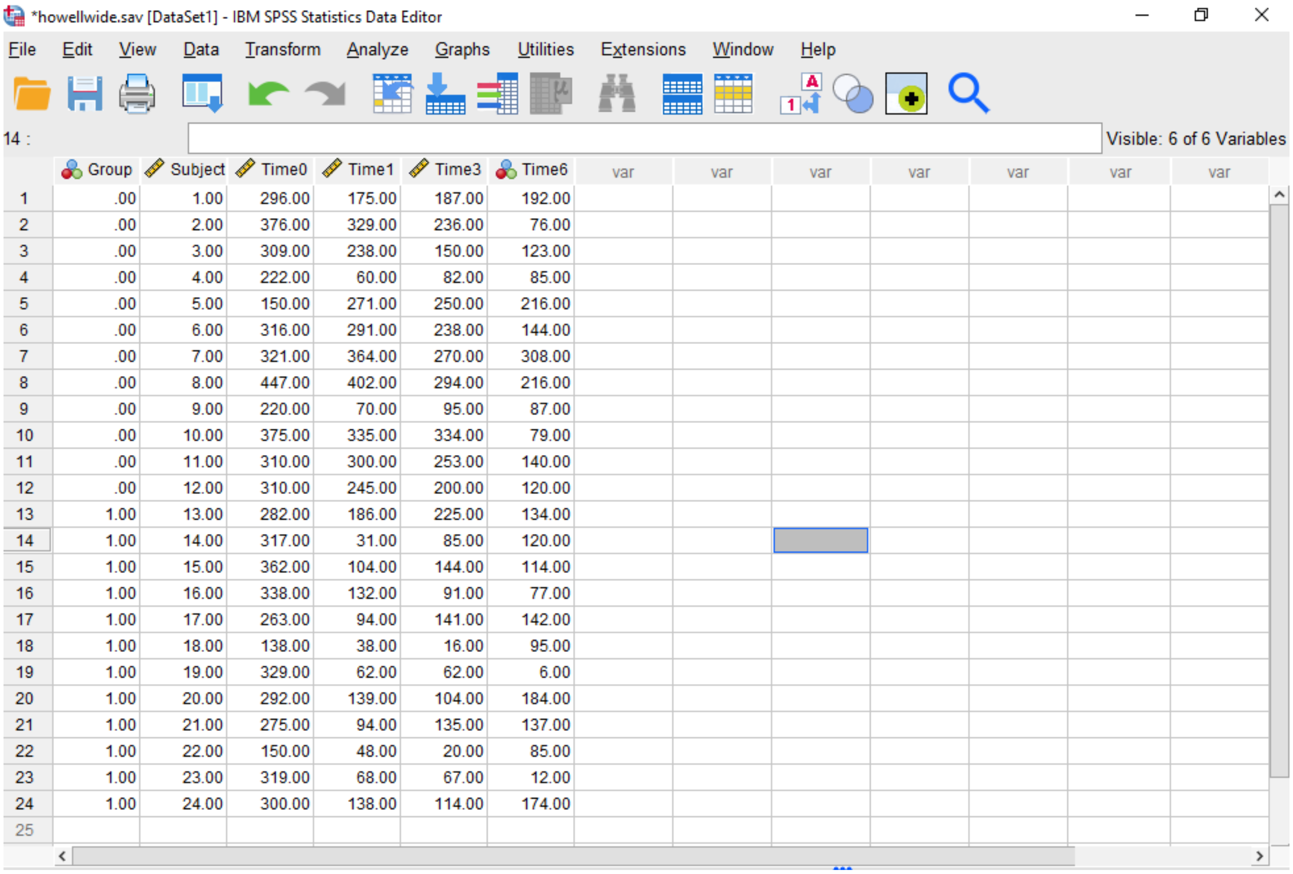
assumptional and practical limitations, and because there are better alternatives (e. g., Gibbons et al., 2010; Vickers, 2005). The glaring limitation to the mixed model ANOVA is that participants with any missing data are discarded (Lix & Keselman, 2010), unless some sort of alternative strategy (e. g., imputation) is adopted. Intention-to-treat analysis (see White et al., 2011) assumes all participants are retained in the analysis regardless of dropout or other factors. Mixed model ANOVA cannot meet that assumption if there are missing data. Furthermore, many authors (e. g., Field, 2018) express concern about meeting the sphericity assumption in a repeated measures ANOVA, specifically that variances and covariances be equal. However, if there are few or no missing data points and a straightforward analysis is an important consideration, then there may be a role for the mixed model ANOVA approach.

Howell (n.d. 2010) presents a fictitious dataset with and without missing values. The missing values dataset has nine data points missing over seven participants — those seven participants would be lost by performing a mixed model ANOVA, reducing the sample size from 24 to 17. However, our focus is on the dataset without missing values (see Supplemental File). Howell was spurred to create the dataset because of a question asked of him by a Swedish researcher with missing data from a randomized clinical trial of two treatment groups and measures at four time periods (pre, post, three months follow-up, and six months follow-up). Howell generated data with one treatment group, one control group, 12 participants per group, and each participant measured four times (pretest, one month, three months, and six months). One of the characteristics of the data is that while the control group declined over time, some participants more so than others, the treatment group declined substantially more than the control group (see Figure 1). Figure 1 was created using dcousin3.shinyapps.io/superbshiny/ (see O'Brien & Cousineau, 2014).

Group has been recoded so that 0 represents control participants and 1 represents treatment participants (Howell coded it 1 for control and 2 for treatment). The SPSS data file is presented in Figure 2 in the standard wide form. Each participant contributes one line of data. The syntax to reproduce the mixed model ANOVA analysis using SPSS GLM appears below. Polynomial is the default contrast in SPSS that tests for a polynomial trend in the data versus a profile contrast that compares means in pairwise fashion or a reference contrast that compares means to a specified reference group.

```
GLM Time0 Time1 Time3 Time6 BY Group
  /WSFACTOR=Time 4 Polynomial
```

Figure 2 ■ Howell's Data in SPSS Wide Form


*howellwide.sav [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Extensions
Window
Help
14 :
Visible: 6 of 6 Variables

	Group	Subject	Time0	Time1	Time3	Time6	var	var	var	var	var	var	var
1	.00	1.00	296.00	175.00	187.00	192.00							
2	.00	2.00	376.00	329.00	236.00	76.00							
3	.00	3.00	309.00	238.00	150.00	123.00							
4	.00	4.00	222.00	60.00	82.00	85.00							
5	.00	5.00	150.00	271.00	250.00	216.00							
6	.00	6.00	316.00	291.00	238.00	144.00							
7	.00	7.00	321.00	364.00	270.00	308.00							
8	.00	8.00	447.00	402.00	294.00	216.00							
9	.00	9.00	220.00	70.00	95.00	87.00							
10	.00	10.00	375.00	335.00	334.00	79.00							
11	.00	11.00	310.00	300.00	253.00	140.00							
12	.00	12.00	310.00	245.00	200.00	120.00							
13	1.00	13.00	282.00	186.00	225.00	134.00							
14	1.00	14.00	317.00	31.00	85.00	120.00							
15	1.00	15.00	362.00	104.00	144.00	114.00							
16	1.00	16.00	338.00	132.00	91.00	77.00							
17	1.00	17.00	263.00	94.00	141.00	142.00							
18	1.00	18.00	138.00	38.00	16.00	95.00							
19	1.00	19.00	329.00	62.00	62.00	6.00							
20	1.00	20.00	292.00	139.00	104.00	184.00							
21	1.00	21.00	275.00	94.00	135.00	137.00							
22	1.00	22.00	150.00	48.00	20.00	85.00							
23	1.00	23.00	319.00	68.00	67.00	12.00							
24	1.00	24.00	300.00	138.00	114.00	174.00							
25													

```
/METHOD=SSTYPE(3)
/PLOT=PROFILE(Time*Group)
/EMMEANS=TABLES(OVERALL)
/EMMEANS=TABLES(Group)
/EMMEANS=TABLES(Time)
/EMMEANS=TABLES(Group*Time)
/PRINT=DESCRIPTIVE ETASQ
/WSDESIGN=Time
/DESIGN=Group.
```

Output 1 at the end presents abbreviated mixed model ANOVA output. Looking at the *sphericity assumed* results in the output, there is a statistically significant *Time* main effect, $F(3,66) = 45.13$, $p < .001$, $\eta_p^2 = .672$, and a statistically significant *Group* main effect, $F(1,22) = 13.71$, $p < .001$, $\eta_p^2 = .384$. Partial eta-squared $\left(\eta_p^2\right)$ is a measure of effect size in which the effects of other independent variables are set aside (i.e., removed from the calculation of the total sums of squared; see Cohen, 1973). These effects are qualified by a statistically significant *Group* by *Time* interaction, $F(3,66) = 9.01$, $p < .001$, $\eta_p^2 = .291$. Thus, we have statistically significant effects that explain a substantial proportion of the variability in the scores.

The mixed model ANOVA assumes that the variances of the pairwise differences between each treatment are equal (sphericity), and equal across groups (multisample sphericity). For this presentation, we will adopt the more straightforward (and strict) compound symmetry assumption that variances and covariances are equal. In other words, there is one variance and one covariance (the same variance and the same covariance) for our four measurements across both groups.

The correlations between scores by group (see Table 1) suggests the compound symmetry assumption has been violated in so far as the correlations (the covariances standardized) are very different within and across groups. Looking at the treatment group (Group 1), the correlations range from $r = .784$ for Time 1 to Time 3 to $r = -.075$ for Time 0 to Time 6. Lix and Keselman (2010) regard the compound symmetry assumption to be important in the-

Table 1 ■ Intercorrelations for Time Disaggregated by Group

Variable	Time0	Time1	Time3	Time6
Time0	–	.375	.482	$-.075$
Time1	.684∗	–	.784∗∗	.493
Time3	.573	.919∗∗	–	.484
Time6	.023	.448	.382	–

Note. The results for Group 0 (Control; $n = 12$; Greenhouse-Geisser ϵ = .735) are shown below the diagonal. The results for Group 1 (Treatment; $n = 12$; Greenhouse-Geisser ϵ = .615) are shown above the diagonal. * $p < .05$, ** $p < .01$ (two-tailed)

ory yet rarely met in practice. Popular alternatives are the Greenhouse-Geisser and Huynh-Feldt corrections to the degrees of freedom in ANOVA (Note that the Greenhouse-Geisser ϵ values, a measure of sphericity that ranges from 0, complete lack of sphericity, to 1, perfect sphericity, are added to Table 1 for each group). The safest route if one is employing the mixed model ANOVA approach is to use the Greenhouse-Geisser or Huynh-Feldt tests.

Anything larger than a 2 × 2 design produces omnibus test results that demand further analysis. Jaccard and Guilamo-Ramos (2002b, 2002a) discuss how to break down the results from an omnibus mixed model ANOVA. Their suggestion, citing Keselman and Keselman (1993), is to judiciously conduct simple main effect contrasts (i.e., exploring the effect of one variable at each level of the second variable). These simple main effect contrasts can be conducted in one of two ways: 1) run dependent *t*-tests (or equivalently repeated measures ANOVAs) separately for each group; or 2) compare the groups with independent samples t tests at each time point.

For the sake of argument, say that a simple main effect contrast is statistically significant for the treatment group but the same simple main effect contrast is not statistically significant for the control group. On that basis, can you conclude that the treatment group is superior to the control group? According to Nieuwenhuis et al. (2011), concluding yes is a frequently made mistake; "Although superficially compelling… the difference between significant and not significant need not itself be statistically significant" (p. 1105).

What is needed is a test to determine if the *difference* between two simple main effects is statistically significant. Jaccard and Guilamo-Ramos (2002b) recommend researchers perform interaction contrasts. In our example, we would partition the 2 (*Group*) × 4 (*Time*) interaction into a series of 2 × 2 interaction contrasts. The easiest way to do so is to select, for example, the pre-test and the six month periods for the treatment and control groups (omitting the one month and three month periods) and then rerun the mixed model ANOVA. Since the recommendation in repeated measures ANOVA is to use an error term derived only from the levels being compared (Howell, n.d.), this is a straightforward and logical method. Since multiple interaction contrasts will need to be conducted, some may raise concerns regarding multiplicity control (e. g., Maxwell & Delaney, 2004) . However, note that the focus of these interaction contrasts is to understand the nature of the significant interaction, and that researchers should focus on the effect sizes more than statistical significance testing (Cribbie, 2017). For our data, the interaction contrast of *Group* (treatment, control) by *Time* (pre-test vs. six months) was not statistically significant, $F(1, 22) = .20$, $p = .66$, $\eta_p^2 = .009$, suggesting the impact of our fictional treatment relative to the control did not persist over half a year. However, the corresponding interaction contrast for pre-test vs. one month was statistically significant and explained a substantial proportion of the variability in our treatment, $F(1, 22) = 21.14$, $p < .001$, $\eta_p^2 = .490$. We recommend running all interaction contrasts of interest to understand the nature of the omnibus interaction.

The MLM Approach

MLM has a number of advantages over the mixed model ANOVA for analyzing TCPPF designs (see Gibbons et al., 2010; Hesser, 2015; Rausch et al., 2003). One advantage is how MLM handles missing data. If a participant misses even one time period in a traditional mixed model ANOVA setting, the default with SPSS, and most other software packages, is to discard all their data (i.e., listwise deletion). Although there are methods to address data missing at random (or completely at random), such as multiple imputation (e. g., imputing the missing value(s) multiple times via a stochastic regression model on the available data, and averaging the results obtained from these complete datasets), MLM uses maximum likelihood (ML) estimation to reach a solution that retains all participants regardless of missing data without the need for imputation. Using all available data is an efficient strategy for estimating model parameters (e. g., variances at each time point). MLM can also flexibly handle continuous and, when time is treated continuously, irregularly spaced measurement periods. The MLM can also obviously handle categorical representations of time; this is how time is conceptualized here due to the nature of the measurements.

A second advantage of the MLM approach is that one can anticipate that participants' scores on the same measure are correlated. The mixed model ANOVA provides little flexibility in terms of modeling the structure of the within person variance-covariance matrix (Chan, 2004), whereas MLM allows for the specification of a variety of variance/covariance structures for correlated measures and their associated errors.

A third advantage of MLM is that it provides the opportunity to evaluate random effect factors. Fixed effect factors assume that a single parameter holds for all participants/groups. For example, if we assume that the intercept is *fixed*, then the value of the outcome is assumed to be constant across all units within the same group. In contrast, random effect factors assume that any group/individual's parameter can vary [e. g., level (random intercept), growth trajectory (random slope), or both]. Random effects can speak to variation between and within subjects. Say we are modeling the change in a single group of participants over two time points. The model could have a random intercept and fixed slope, where individuals can vary on the level of the outcome, but each individual would have the same slope (i.e., parallel lines). Or the model could have a random intercept and random slope where individual levels (e. g., starting points) and slopes can vary (i.e., non-parallel lines with varying starting points). Mixed model ANOVAs are fixed effect models with the exception of random residuals. MLM allows for fixed and random effect models or combinations of fixed and random effects. In addition to incorporating random variation into the analysis, another benefit of incorporating random effects is the ability to estimate the variances associated with the varying effect estimates (e. g., by how much do the intercepts differ across participants?).

As a starting point, we will replicate the mixed model ANOVA in SPSS `MIXED`, the dedicated program for running MLM in SPSS. The first step is to transform Howell's data from wide form to long form. That can be done through the following syntax:

```
VARSTOCASES
  /MAKE DV FROM Time0 Time1 Time3 Time6
  /INDEX=Time(4)
  /KEEP=Group Subject
  /NULL=KEEP.
```

One should inspect the resulting dataset to make sure the transformation is correct, especially important if there are missing data. Figure 3 presents a screenshot of a selection of the long form data for the Howell data. For each participant, there are four rows of data in the long form (one for each time point) rather than a single row of data in the wide form. Participants are designated by "Subject". We have a variable "Time" to capture whether a score was generated at pre-test (Time = 0), at one month (Time = 1), three months (Time = 3) or six months (Time = 6). Recall that time is treated as categorical, not continuous. "Group" indicates if a participant is associated with the treatment group (Group = 1) or the control group (Group = 0). We have one outcome variable labelled as "dv" rather than four outcome variables each associated with each measurement time.

It is a good idea to do a plot of the data before plunging into MLM. The following syntax will produce a line graph for our two groups. See Figure 4.

```
GRAPH
  /LINE(MULTIPLE)=MEAN(dv) BY Time BY Subject
  /PANEL COLVAR=Group COLOP=CROSS.
```

One observes (a) the treatment group (Group = 1) starts lower than the control group (Group = 0), (b) most of the treatment group starts at about the same spot (near the mean of the outcome variable), while the control group is more spread out in their starting scores, (c) almost all participants get better over time regardless of group, (d) many of the treatment group scores rise between Time = 3 and Time = 6, which is not true of the control group, and (e) the treatment group, for the most part, has lower scores at the last time point than the control group.

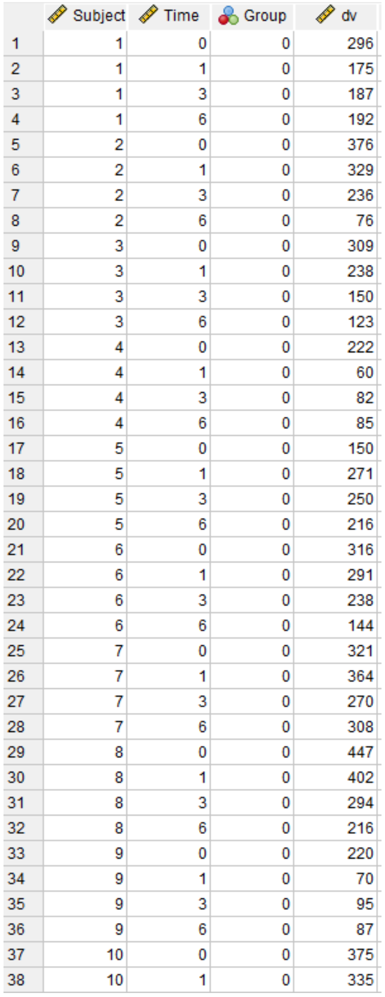
Marginal Model. What many sources, including Howell (n.d.), first calculate, but do not label as such, is the marginal model. The marginal model takes the form of a MLM, but has no random effects (see Heagerty & Zeger, 2000). The marginal model is the equivalent of the mixed model ANOVA except the former uses ML estimation. Selecting a marginal model over MLM is appropriate when you measure the outcome variable only a few times, when there are no higher level clusters (e. g., participants counter-balanced), when your interest is in mean differences across groups rather than change over time, which is better captured by MLM, when MLM fails to converge, and when you are interested in comparing different covariance structures but not random effects

The following syntax replicates the mixed model ANOVA in SPSS MIXED for the marginal model:

```
MIXED dv BY Group Time
  /FIXED=Group Time Group*Time | SSTYPE(3)
  /METHOD=REML
  /PRINT=SOLUTION TESTCOV R
  /REPEATED=Time | SUBJECT(Subject) COVTYPE(CS).
```

The first line of syntax tells SPSS to access SPSS `MIXED`. The outcome variable is "dv" and the independent variables are "Group" and "Time". The outcome variable must be listed first. BY appears innocuous but it is the first of many complexities in SPSS MIXED. BY identifies categori-

Figure 3 ■ SPSS Data in Long Format



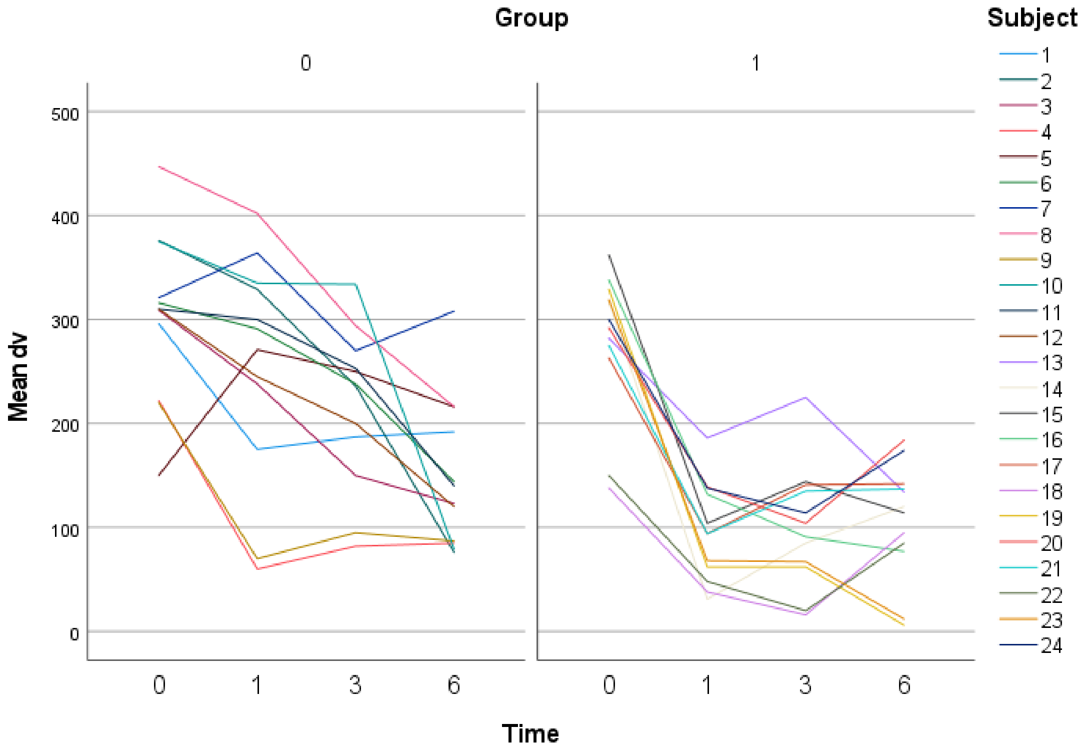
	Subject	Time	Group	dv
1	1	0	0	296
2	1	1	0	175
3	1	3	0	187
4	1	6	0	192
5	2	0	0	376
6	2	1	0	329
7	2	3	0	236
8	2	6	0	76
9	3	0	0	309
10	3	1	0	238
11	3	3	0	150
12	3	6	0	123
13	4	0	0	222
14	4	1	0	60
15	4	3	0	82
16	4	6	0	85
17	5	0	0	150
18	5	1	0	271
19	5	3	0	250
20	5	6	0	216
21	6	0	0	316
22	6	1	0	291
23	6	3	0	238
24	6	6	0	144
25	7	0	0	321
26	7	1	0	364
27	7	3	0	270
28	7	6	0	308
29	8	0	0	447
30	8	1	0	402
31	8	3	0	294
32	8	6	0	216
33	9	0	0	220
34	9	1	0	70
35	9	3	0	95
36	9	6	0	87
37	10	0	0	375
38	10	1	0	335

cal predictors whereas WITH lists covariates that are to be treated as continuous. There can be one BY and one WITH command in the same `MIXED` statement (e. g., `BY Group WITH Time`). In our example, the between-subject variable *Group* is categorical so it should be entered with a `BY` statement. Similarly, the within-subject variable *Time* is what Howell (n.d.) refers to as an ordered categorical variable (since it was not measured at evenly spaced intervals and it contains an intervention between two of the time points) so it should also be entered with a BY statement. If we had several time points over which we would expect linear change, then treating *Time* as continuous would allow us to estimate the per-unit-of-time change. The F values associated with *Time* and the *Group* by *Time* interaction vary widely depending on if *Time* is entered using `BY` or `WITH` (which is expected since if we consider time to be continuous we would be looking for a linear relationship over time, instead of exploring differences among all levels of the categorical variable). The bottom line is that our mixed model ANOVA results replicate only if Time and Group are entered via BY.

The next line is `/FIXED`. What follows is a list of fixed effects, specifically *Group*, *Time*, and *Group* × *Time*. `SSTYPE(3)` is how sums of squares will be partitioned, with this default being the traditional Type III sum of squares (this selection may or may not be appropriate; see Smith & Cribbie, 2014).

`/METHOD` is the estimation method: ML or Restricted Maximum Likelihood (REML). ML is better if you are examining fixed effects (as per a strictly traditional analysis) or comparing models that differ in their fixed effects. Alternatively, REML is better if you are examining random effects

Figure 4 ■ Line plot.



or if you are comparing random effect models (McNeish, 2017). However, REML and ML frequently give similar results (Everitt & Pickles, 2004).

`/PRINT SOLUTION` provides estimates of values associated with our `FIXED` command. `TESTCOV` is an option to provide Wald test statistics and confidence intervals for variance estimates. R prints the variance-covariance matrix for the residuals. Examining the variance-covariance matrix is essential for interpreting the results from the variance estimates.

The final and critical line of syntax to replicate the mixed model ANOVA results in SPSS `MIXED` is `/REPEATED`. This line provides variances and covariances associated with repeated score (level 1) residuals and maps the relationships between those residuals based on their presumed covariance structures. `Time|SUBJECT(Subject)` indicates that the time points are nested within the subject IDs.

SPSS `MIXED` provides for a number of different possible covariance structures in the `/REPEATED` statement. Authors of sources on SPSS `MIXED` (and researchers) frequently choose covariance structures without specifying or explaining their choice. Like (Howell, n.d.), we have chosen Compound Symmetry (`CS`) because that is the equivalent of assuming equal variances and covariances as per the mixed model ANOVA. An alternative in SPSS `MIXED` that was also chosen by Howell is Unstructured (`UN`) which allows each time period to have its own variance and each pair of time periods their own covariance. UCLA Statistical Consulting Group (n.d.) wrote “If the compound symmetric covariance is overly simple, the unstructured covariance seems overly complex” (para 27). Howell (n.d.) was even more colorful in his description of UN: “Put another way, with the unstructured solution we threw up our hands and said to the program ‘You figure it out! We don’t know what’s going on’” (para 37). Again, if we adopted UN with four measures we would need to estimate four variances and six covariances, a substantial number of parameters given our small sample size (see Chan, 2004). With CS, there are only two parameters.

The marginal model output is presented in Output 2. *Model Dimension* reports number of levels (not to be confused with levels in multilevel modeling) for each predictor in the model, with one for the intercept (since it is continuous), two for *Group* (two groups), four for *Time* (four time points), and eight for *Group* × *Time* (2 groups × 4 time points). There is one parameter for the intercept, one parameter for *Group* (number of groups – 1 because of dummy coding), three parameters for *Time* (number of time points – 1 because of dummy coding), and three parameters for *Group* × *Time* [(number of groups – 1) × (number of time points - 1)]. There are two repeated parameters for *Time*, specifically one variance and one covariance because of our choice of compound symmetry that assumes the same variance and same covariance for all levels of *Time*. *Information Criteria* values permit comparing models with different covariance structures, with lower values indicating better model fit. For compound symmetry, the log-likelihood value is 1000.805, Akaike's Information Criterion (AIC) is 1004.805, and Schwarz's Bayesian Criterion (BIC) is 1009.76. Both AIC and BIC penalize for model complexity; BIC penalizes more than AIC. Other covariance structures will provide higher or lower values for log-likelihood, AIC, and BIC.

Type III Tests of Fixed Effects are identical to our mixed model ANOVA results, specifically statistically significant main effects for group and time, and a statistically significant interaction of group by time. Turning to *Estimates of Fixed Effects*, the intercept is 106.67, the mean for Treatment at Time 6. Time 6 for Treatment is our intercept because of how we coded our variables in the dataset. SPSS defaults to the last category (i.e., Time = 6, Group = 1) as the reference category.

The reference group is Treatment (Group = 1) and the effect of the Control (Group = 0) is 42.17 (see Output 2 at the end). Thus, the Control is scoring 42.17 higher than the Treatment at *Time* 6. Thus, 106.67 plus 42.17 is equal to 148.84 or the mean for Control at *Time* 6. The next estimate is 173.75, labeled '[Time = 0]', which represents the effect of time from *Time* 6 to *Time* 0 for Treatment; if you add 173.75 plus the intercept of 106.67, that equals to 280.42 or the mean for Treatment at *Time* 0. Next is -12.17, labeled '[Time = 1], the effect of *Time* from *Time* 6 to *Time* 1. The intercept of 106.67 plus -12.17 is equal to 94.5, which is the mean for Treatment at *Time* 1. Similarly, the intercept of 106.67 plus -6.33 is equal to 100.34 or the mean for Treatment at *Time* 3. And the zero that follows — the intercept is 106.67 and 106.67 (plus 0) is the mean for Treatment at *Time* 6. What about the -18.25, 120, and 73.25? Those relate to the means of the interaction contrasts between *Group* and *Time*. For example, -18.25 is equal to

$$\begin{aligned}&(\bar{X}_{Time\ 0,\ Treatment} - \bar{X}_{Time\ 0,\ Control})\\&-(\bar{X}_{Time\ 6,\ Treatment} - \bar{X}_{Time\ 6,\ Control})\\&=(304.33-280.42)-(148.83-106.67)\,.\end{aligned}$$

In other words, this represents the difference between the treatment effect at Time 0 and the treatment effect at Time 1. These values are important because they highlight the interaction contrasts that drive the omnibus interaction effect. In this example, we see that the *Group* × *Time* (0,6) interaction contrast contributes little to the interaction, but the *Group* × *Time* (1,6) interaction contrast and *Group* × *Time* (3,6) interaction contrast both contribute to the interaction. Contrast tests for interactions are discussed below. SPSS MIXED does not provide measures of effect size given the complexity of calculating effect sizes with multiple error terms, however see Rights and Sterba (2019) for possibilities.

Next we see *Estimates of Covariance Parameters*. These values are best understood in relation to the Residual Covariance (R) Matrix. Careful, though — the Estimates of Covariance Parameters table is NOT listing the variance and the covariance, per se. According to Littell et al. (2006; see also Littell et al., 2000), the *CS Covariance* value of 2539.361 is the covariance between two scores for the same subject and is equivalent to a between-subjects variance component that Littell et al. (2006) label as σ_s^2. On the other hand, the *CS diagonal offset* value of 2760.622 is an estimate of a residual variance component or the variance conditional on the participant that Littell et al. label as σ_B^2. That 2760.622 is the same value as the mean squared error found for *Sphericity Assumed* in the mixed model ANOVA (see Output 1), given we have a balanced design and no missing data. The total variance or σ^2 according to Littell et al. is the sum of between-subjects variance (covariance) and the residual variance, that is $\sigma^2 = \sigma_s^2 + \sigma_B^2$ or 5299.98 =2539.361 + 2760.622. In the R matrix (Output 2), you see that the diagonal is the variance (5299.98), the off-diagonal is the covariance (2539.361).

An additional option is /TEST. This option could be used to produce contrast tests. We might have tested an interaction contrast between scores at Time 0 and Time 1 by *Group* (treatment and control) via:

```
/TEST(0) Group*Time 1 -1 0 0 -1 1 0 0.
```

As noted above, other covariance structures could be selected. For example, we could choose an unstructured or UN covariance structure by replacing CS with UN as Howell (n.d.) attempted. See Output 3 for selected output. *Tests of Fixed Effects* and *Estimates of Fixed Effects* are identical between CS and UN so are not reported. Note that for *Model Dimensions*, the number of *Repeated Effects Time* parameters is 10 (versus two for CS), specifically four variances and six covariances. The log-likelihood is 975.37 (vs. 1000.80 for CS) and AIC is 995.37 (vs 1004.80 for CS) so both are better for UN. However, BIC which penalizes for model complexity is 1020.15 for UN (vs. 1009.76 for CS). Examination of other covariance structures reveals worse model fit compared to UN or CS. For the UN model, the *Estimates of Covariance Parameters* is best understood relative to the *Residual Covariance R Matrix*. UN (1,1), UN (2,2), UN (3,3), and UN (4,4) values are the variances in the residual covariance matrix (i.e., σ_1^2, σ_2^2, σ_3^2, σ_4^2). In that vein, UN (2,1) or 3535.58 is the residual covariance (i.e., σ_{21}) between Time 0 and Time 1 (recall that both variances and covariances can vary across time points in the UN covariance structure). There is little covariance between Time 0 and Time 6 ($\sigma_{41} = -80.2575$) relative to the covariances between other time periods.

Autoregressive or AR1 is an alternative to CS and UN (see Output 4). As stated by Howell (n.d.), "[AR] assumes that correlations between any two times depend on both the correlation at the previous time and an error component. To put that differently, your score at time 3 depends on your score at time 2 and error" (para 37). AR1 produces a log likelihood of 991.55 (so worse than for UN) and AIC of 995.55 for the AR1 structure (the same as for UN). The AR1 structure estimates one variance (AR1 diagonal = 5190.42, SE = 1005.51) and one correlation (AR1 rho = .611, SE = .083). The diagonals of the R matrix are $\sigma^2 = 5190.42$, the off-diagonals of the R matrix being $\sigma^2 \times \rho$ ($5190.4^2 \times .611 = 3172.215$), $\sigma^2 \times \rho^2$ ($5190.4^2 \times .611^2 = 1938.754$), and $\sigma^2 \times \rho^3$ ($5190.4^2 \times .611^3 = 1184.903$).

Random Intercept Model. So far we have focused on fixed effects in MLM because that is all the mixed model ANOVA and the marginal model are capable of examining. Fixed effect factors are appropriate if it is plausible that all individuals/groups have the same parameter (e. g., intercept). In contrast, if a factor (e. g., the intercept) varies across participants and thus would be better represented by a distribution rather than a point estimate, then this factor could be specified as a random factor. As stated by McNeish (2017), "the random effects… capture the difference between aspects of the [individual/group] specific intercept and slopes and the overall regression line formed by the fixed effects" (p. 663).

Adding random effects adds complexities to the analysis. Chan (2004) cautions that "[f]or the extension of the fixed effects to a mixed effect model (having both fixed and random effects), it would be most appropriate to seek the assistance of a biostatistician!" (p. 460). However, Judd et al. (2012), among others, encourage greater use of random effect models by psychological researchers.

In the marginal model, we modeled *Time* using the REPEATED statement. To account for dependencies between scores using random effects in SPSS MIXED, we swap out the /REPEATED statement for a /RANDOM statement that lists predictors of random effects in the models. A RANDOM statement serves to specify random effects through the G matrix while a REPEATED statement speaks to the structure of the R matrix (see Littell et al., 2006).

The random intercept model allows each individual's intercept (pre-test level, which in SPSS is by default the last level) to vary from the average for the individual's condition (i.e., treatment or control group). The goal is to determine the variability associated with those individual scores and determine if that variability is substantial. One needs only to change the last line of the syntax from the marginal model to run the random intercept model, specifically:

```
/RANDOM=Intercept | SUBJECT(Subject)
 COVTYPE(VC).
```

The /RANDOM statement specifies a trajectory for each "Subject" or participant within each group, all with the same slope but with a random intercept specific to each participant. SUBJECT() lists sampling units. In our example, participants are randomly assigned to treatment or control groups at Level 2 while their repeated measurements are at Level 1. COVTYPE(VC) or Variance Components is the covariance structure suggested by Field (2018). Other covariance structures are converted to a VC or unstructured covariance structure given our random effect and G matrix has only one parameter (i.e., the random intercept). One should add G after R in the /PRINT command to obtain the G matrix, the matrix associated with random intercepts (and if applicable, random slopes).

The resulting output for the random intercept model is identical to the output from the marginal model with a couple of exceptions (see Output 5). Looking at *Model Dimension*, there is one parameter associated with the random intercept and one parameter associated with residual error. Fixed effects are unchanged. Turning to the *Estimates of the Covariance Parameters*, the variances around the fixed effects, the value 2760.62 is now labelled *Residual* or the level 1 residual variance. Recall for the mixed model ANOVA this variance was labeled the mean squared error and for the marginal model this variance was labeled the *CS diagonal offset*. These three models are referring to the same residual variance using different terms. The R matrix has only the one value of 2760.62 because the VC covariance structure has the same variance for all four time periods and zeros for the covariances (having defaulted to a Scaled Identity matrix that is assigned to each random effect). The value of 2539.36 is now labelled the *Intercept [subject=Subj] Variance* and is the variance estimate (τ^2) of the intercept (Field, 2018). The G matrix value reflects the variance associated with the random effect of the intercept of 2539.36. Both the residual and intercept variance estimates are statistically significant, indicating that there is variability remaining that could be explained by adding time-varying predictors (to explain differences over time) or time-invariant predictors (to explain variability in the intercepts).

The random intercept model adds the G and R matrices after the G matrix is converted to a 4x4 matrix (reflecting our four measurement periods). The resulting variance-covariance matrix is equivalent to the compound symmetry (CS) matrix of constant variances and constant covariances, hence why we are getting the same results from the marginal model and the random intercept model (albeit with different labels for *Estimates of the Covariance Parameters*). And we also get the same results from a repeated measures ANOVA model given we have a balanced design and no missing data.

Given that the output from the marginal model and the random intercept model frequently do not differ except in their labeling in SPSS MIXED, comparing those two models is of no benefit. What is of benefit is to compare the random intercept model to a simpler form of that same model, the empty model (Output 6). The empty model is run as per the random intercept model in SPSS MIXED except that /FIXED= has nothing following the equal sign. Doing so allows one to calculate and compare the intraclass correlation coefficient (ICC) values for the two models or $\tau^2/(\tau^2 + \sigma^2)$. Note that τ^2 represents in this instance the intercept (between subject) variance and σ^2 represents the residual (individual subject) variance. The random intercept model with fixed effects produced an ICC of .479 [2539.36 / (2539.36+2760.62)]. That compares to an ICC of .244 for an empty model with no fixed effects or 2824.89 / (2824.89+8759.14). For the empty model, note the substantial within subjects/residual variance of $\sigma^2 = 8759.14$ relative to the variance between subjects or attributable to the intercept of $\tau^2 = 2824.89$. Also note the substantial relative drop in within subjects or residual variance from $\sigma^2 = 8759.14$ to $\sigma^2 = 2760.62$ as we move from the empty model to the random intercept model (i.e., by considering the individual level intercepts we are much better able to predict the scores of the participants on the outcome variable). It is noteworthy that the random intercepts are not tied to a specific reference group but instead relate to overall between participant differences in the level of the outcome.

Random Slope Model. The random slope model (that also includes a random intercept) evaluates the assumption that each participant's slope or trajectory of change varies from the average for the participant's condition (i.e., treatment or control group). It might be important for the researcher to determine the variability associated with those individual trajectories and determine if that variability is substantial. The random slope model can be beneficial for determining if there are time-varying covariates that might help explain variability in slopes over time. As with the random intercept model, the random slope model improves the operation of our fixed effect model.

Howell (n.d.) proceeded to perform a random slope model, albeit noting his uncertainty as how to proceed. When time is treated categorically (as we have done in this example), it is not possible to estimate variability in slopes between each of the time points. When we request random slopes, there are three slope variances for our four time periods for each participant. Time 6 is the reference by default in SPSS; thus, there is a slope variance for Time 0 vs. Time 6, there is a slope variance for Time 1 vs. Time 6, and there is a slope variance for Time 3 vs. Time 6. For example, the Time 0 vs Time 6 random slope would estimate the variability in slopes across these two time points. Thus, we are requesting a random intercept value, three random slope values, and a residual variance value (for the random effects). When treating time categorically, it can be impossible to estimate random slopes because we are not providing enough information (e. g., variances, covariances from the time points) to be able to estimate the individual slopes (and thus the individual slope variances) in addition to the other model parameters (e. g., intercept variances, residual variance). This difficulty might seem confusing since we know that it is easy to calculate the variances of the slopes via a simple difference score model (i.e., calculate the variance of Time 1 – Time 0, the variance of Time 3 – Time 0, etc.). However, what is explicit in these models is that the residual is zero; the mixed model that we have been running estimates the residual and it is not possible to fix it to zero in SPSS. Some software programs, for example the structural equation modeling program AMOS, allow users to fix residual parameters to zero; doing so would allow us to estimate the model as long as we are comfortable assuming that the residuals are all zero.

Only for a different design in which each individual at each time period was associated with multiple measurements would it be possible to estimate a random slope model (Singmann & Kellen, 2019). Imagine a study looking at the effect of cognitive behavioral therapy on depression; individuals were measured across four waves (stages of therapy), and at each wave there were multiple measurements of depression (e. g., different depression scales; measurements before, during and after the counselling session). Thus, in most mixed model design cases, we should focus our attention in evaluating on either the random intercept model or the marginal model.

Presentation of Results

Below we present the results for the mixed model ANOVA, marginal model, and random intercept model. These results should be supplemented by appropriate tables and figures. Correlations between scores on the variables should also be reported. To summarize, if we have a group by time design with all participants measured at the same points in time, if we have no missing data, if we are satisfied with a compound symmetry covariance structure, and if we want a relatively simple analysis, then the mixed model ANOVA is sufficient. If there are missing data, if we wish to explore other covariance structures, if time periods vary between participants, and if we are willing to tackle more complex analyses, then MLM is appropriate. The marginal model and the random intercept model produce similar results, but if we wish to explore random variables and more complex designs, then random intercept (and slope) models might be investigated. Otherwise, a marginal model buys one the advantages of MLM without the added complexity.

What might our results look like as a mixed model ANOVA? A mixed model ANOVA was conducted on the data from the 24 participants from Howell (n.d.). Group (treatment or control) was a between-subject factor. Time (four time periods) was the within-subject factor. There was a main effect for Group, $F(1, 22) = 13.714$, $p < .001$, $\eta_p^2 = .384$, and a main effect for Time, $F(3, 66) = 45.135$, $p < .001$, $\eta_p^2 = .672$. The main effects were qualified by a statistically significant interaction of Group by Time, $F(3, 66) = 9.014, p < .007, \eta_p^2 = .291$. According to cutoffs (e. g., Cohen, 1988), partial eta-squared greater than 0.14 indicates a large effect (these cutoffs actually apply to eta-squared but have been transferred to partial eta-squared). Cell means are plotted in Figure 1. Scores diminished over time, more so for the treatment group. Interaction contrasts (group by time) were statistically significant for pre-test (Time 0) and one month (Time 1), $F(1, 22) = 21.141$, $p < .001$, $\eta_p^2 = .490$, pre-test (Time 0) and three months (Time 3), $F(1, 22) = 10.586$, $p < .004$, $\eta_p^2 = .325$, one month (Time 1) and three months (Time 3), $F(1, 22) = 7.247$, $p < .013$, $\eta_p^2 = .248$, one month (Time 1) and six months (Time 6), $F(1, 22) = 13.738, p < .001, \eta_p^2 = .384$, three months (Time 3) and six months (Time 6), $F(1, 22) = 6.372, p < .019, \eta_p^2 = .225$, but not pre-test (Time 0) and six months (Time 6), $F(1, 22) = .204, p < .656, \eta_p^2 = .009$.

How about marginal multilevel model results? A marginal multilevel model was conducted on the data from the 24 participants from Howell (n.d.). Group (treatment or control) was a between-subject factor. Time (four time periods) was the within-subject factor. The compound symmetry covariance structure was selected. There was a main effect for Group, $F(1, 22) = 13.714, p < .001$, and a main effect for Time, $F(3, 66) = 45.135$, $p < .001$. Again, the main effects were qualified by an interaction of Group by Time, $F(3, 66) = 9.014, p < .007$. Cell means are plotted in Figure 1. The residual covariance between each (and all) of the time periods is 2539.36 ($SE = 981.119$). The residual variance was 2760.622 ($SE = 480.563$). The total variance is the sum of the residual covariance and residual variance or 5299.983. Statistically significant interaction (group by time) contrasts were found between pre-test (Time 0) and one month (Time 1), $\psi_{01} = -138.25$ ($SE = 30.335$), $t(66) = -4.557$, $p < .001$, 95% CI $[-198.816, -77.684]$, and pre-test (Time 0) and three months (Time 3), $\psi_{03} = -91.500$ ($SE = 30.335$), $t(66) = -3.016$, $p < .004$, 95% CI $[-152.066, -30.934]$, but not between pre-test (Time 0) and six months (Time 6), $\psi_{06} = -18.250$ ($SE = 30.335$), $t(66) = -.602$, $p < .549$, 95% CI $[-78.816, 42.316]$. The residual variance was $\sigma^2 = 2760.62$ ($SE = 480.563$), a value that was statistically significant, Wald $Z = 5.745$, $p < .001$, 95% CI $[1962.600, 3883.131]$. Therefore, there could be potential level 1 predictors that explain a portion of this variance.

Finally, what would the results for a random intercept model look like? A random intercept multilevel model was conducted on the data from the 24 participants from Howell (n.d.). Group (treatment or control) was a between-subject factor. Time (four time periods) was the within-subject factor. The variance components covariance structure was selected. There was a main effect for Group, $F(1, 22) = 13.714$, $p < .001$, and a main effect for Time, $F(3, 66) = 45.135$, $p < .001$. These main effects were qualified by a statistically significant interaction of Group by Time, $F(3, 66) = 9.014$, $p < .007$. Cell means are plotted in Figure 1. Statistically significant interaction contrasts [as above]… The residual variance was [as above]… The random intercept model with fixed effects produced an intraclass correlation coefficient of ICC = .479 compared to an empty random intercept model with no fixed effects of ICC = .244.

Discussion

More than 20 years ago, Overall et al. (1999) wrote of their struggles to understand SAS PROC MIXED, the SAS equivalent to SPSS MIXED. Overall et al. spoke of how "hidden in the general formulation of the mixed model equation are liabilities associated with what is perhaps too great flexibility for defining a model on which inferences about treatment effects are to be based" (p. 190). Like us, they relied upon the literature for guidance on how to analyze TCPPF data, and, like us, they reported conflicting and unclear advice. Their recommendation was that researchers must specify in detail exactly how their analyses were run; "[g]iven the procedure's inherent flexibility, the only way that anyone can evaluate adequacy of reported results is to consider the complete model specification… observed random effects and error correlation matrix" (p. 215). Thus, we conclude with the same recommendation as Overall et al. — that whatever researchers do when they analyze data from a TCPPF design, that data analysis be reported clearly and, for even more clarity, including the syntax used. Since

SPSS is the most popular software program in psychology (Davidson et al., 2019), we believe it makes sense to have a resource available that can assist researchers in implementing and understanding hierarchical/mixed model analyses for TCPPF designs within this software package. We hope readers will find the present tutorial to be useful as they analyze their own TCPPF data.

Authors' note

The authors wish to thank Udi Alter, Nataly Beribisky, Linda Farmus, Naomi Martinez Gutierrez, Danielle Perepelkin, and Julie Ziemer for their helpful comments.

References

Assman, S. F., Pocock, S. J., Enos, L. E., & Kasten, L. E. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*, *355*, 1064–1069. doi: 10.1016/s0140-6736(00)02039-0.

Chan, Y. H. (2004). Biostatistics 301a. repeated measurement analysis (mixed models). *Singapore Medical Journal*, *45*(10), 456–461.

Chester, D. S., & Lasko, E. N. (2021). Construct validation of experimental manipulations in social psychology: Current practices and recommendations for the future. *Perspectives on Psychological Science*, *16*(2), 377–395. doi: 10.1177/1745691620950684.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor anova designs. *Educational and Psychological Measurement*, *33*(1), 107–112. doi: 10.1177/001316447303300111.

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

Counsell, A., & Cribbie, R. A. (2017). Using the errors-in-variables method in two-group pretest-posttest designs. *Methodology*, *13*(1), 1–8. doi: 10.1027/1614-2241/a000122.

Cribbie, R. A. (2017). Multiplicity control, school uniforms, and other perplexing debates. *Canadian Journal of Behavioural Science*, *49*(3), 159–165. doi: 10.1037/cbs0000075.

Cribbie, R. A., & Jamieson, J. (2000). Structural equation models and the regression bias for measuring correlates of change. *Educational and Psychological Measurement*, *60*(6), 893–907. doi: 10.1177/00131640021970970.

Davidson, H., Jabbari, Y., Patton, H., O'Hagan, F., Peters, K., & Cribbie, R. (2019). Statistical software use in canadian university courses: Current trends and future directions. *Teaching of Psychology*, *46*(3), 244–250. doi: 10.1177/0098628319853940.

Eghewale, B. E. (2015). Statistical issues in randomised controlled trials: A narrative synthesis. *Asian Pacific Journal of Tropical Biomedicine*, *5*(5), 354–359. doi: 10.1016/s2221-1691(15)30367-1.

Everitt, B. S., & Pickles, A. (2004). *Statistical aspects of the design and analysis of clinical trials (rev ed.)* Imperial College Press.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed., North American Edition). Sage.

Gibbons, R. D., Hedeker, D., & DuToit, S. (2010). Advances in analysis of longitudinal data. *Annual Review of Clinical Psychology*, *6*(3), 29. doi: 10.1146/annurev.clinpsy.032408.153550.

Guidi, J., Brakmeier, E.-L., Bockting, C. L. H., Coscci, F., Cjipers, P., ..., & Fava, G. A. (2018). Methodological recommendations for trials of psychological interventions. *Psychotherapy and Psychosomatics*, *87*, 276–284. doi: 10.1159/000490574.

Heagerty, P. J., & Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, *15*(1), 1–26. doi: 10.1214/ss/1009212671.

Heck, R. H., Thomas, S. L., & Tabata, L. N. (2014). *Multilevel and longitudinal modeling with IBM SPSS* (2nd ed.). Routledge.

Hesser, H. (2015). Modeling individual differences in randomized experiments using growth models: Recommendations for design, statistical analysis and reporting of results of internet interventions. *Internet Interventions*, *2*, 110–120. doi: 10.1016/j.invent.2015.02.003.

Howell, D. C. (2010). *Statistical methods for psychology* (8th ed.). Nelson.

Howell, D. C. (n.d.). *Overview of mixed models*. Retrieved January 1, 2023, from https://www.uvm.edu/statdhtx/StatPages/Mixed-Models-Repeated/Mixed-Models-Overview.html

Jaccard, J., & Guilamo-Ramos, V. (2002a). Analysis of variance frameworks in clinical child and adolescent psychology: Advanced issues and recommendations. *Journal of Clinical Child and Adolescent Psychology*, *31*(2), 278–294. doi: 10.1207/s15374424jccp3102_13.

Jaccard, J., & Guilamo-Ramos, V. (2002b). Analysis of variance frameworks in clinical child and adolescent psychology: Issues and recommendations. *Journal of Clinical Child and Adolescent Psychology*, *31*(1), 130–146. doi: 10.1207/s15374424jccp3101_15.

Jennings, M., & Cribbie, R. A. (2016). Comparing pre-post change across groups: Guidelines for choosing between difference scores, ancova, and residual change scores. *Journal of Data Science*, *14*, 205–230. doi: 10.6339/JDS.201604_14(2).0002.

Jones, B., & Nachtsheim, C. J. (2009). Split-plot designs: What, why, and how. *Journal of Quality Technology*, *41*(4), 340–361. doi: 10.1080/00224065.2009.11917790.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. doi: 10.1037/a0028347.

Kahan, B. C., Jairath, V., Dore, C. J., & Morris, T. P. (2014). The risks and rewards of covariate adjustment in randomized trials: An assessment of 12 outcomes from 8 studies. *Trials*, *15*(139), 1–7. doi: 10.1186/1745-6215-15-139.

Kendall, P. C., Comer, J. S., & Chow, C. (2013). The randomized controlled trial: Basics and beyond. In J. S. Comer & P. C. Kendall (Eds.), *The oxford handbook of research strategies of clinical psychology* (pp. 40–61). Oxford University Press.

Keselman, H., & Keselman, J. (1993). Analysis of repeated measurements. In L. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 105–146). Dekker.

Knapp, T. R., & Schafer, W. D. (2009). From gain score to ancova f (and vice versa). *Practical Assessment, Research & Evaluation*, *14*(6), 1–7. https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1212&context=pare

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models (2nd ed.)* SAS Institute Inc.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Modeling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, *19*(13), 1793–1819. doi: 10.1002/1097-0258(20000715)19:13<1793::aid-sim482>3.0.co;2-q.

Lix, L. M., & Keselman, H. J. (2010). Analysis of variance: Repeated measures designs. In G. R. Hancock, R. O. Mueller, & L. Stapleton (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 15–27). Taylor & Francis.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective (2nd ed)*. Lawrence Erlbaum Associates Publishers.

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the kenward-roger correction. *Multivariate Behavioral Research*, *52*(5), 661–670. doi: 10.1080/00273171.2017.1344538.

Nieuwenhuis, S., Forstmann, B., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107. doi: 10.1038/nn.2886.

O'Brien, F., & Cousineau, D. (2014). Representing error bars in within-subject designs in typical software packages. *The Quantitative Methods for Psychology*, *10*(1), 56–67. doi: 10.20982/tqmp.10.1.p056.

O'Connell, N. S., Dai, L., Jiang, Y., Speiser, J. L., Ward, R., Wei, W., Carroll, R., & Gebregziabher, M. (2017). Methods for analysis of pre-post data in clinical research: A comparison of five common methods. *Journal of Biomedical Biostatistics*, *8*(1), 1–8. doi: 10.4172/2155-6180.1000334.

Overall, J. E., Anh, C., Shivakumar, C., & Kalburgi, Y. (1999). Problematic formulations of SAS PROC.MIXED models for repeated measures. *Journal of Biopharmacological Statistics*, *9*(1), 189–216. doi: 10.1081/bip-100101008.

Peugh, J. L., & Enders, C. K. (2005). Using the SPSS Mixed procedure to fit cross-sectional and longitudinal models. *Educational and Psychological Measurement*, *65*(5), 717–741. doi: 10.1177/0013164405278558.

Rausch, J. R., Maxwell, S. E., & Kelley, K. (2003). Analytic methods for questions pertaining to a randomized pretest, posttest, follow-up design. *Journal of Clinical Child and Adolescent Psychology*, *32*(3), 467–486. doi: 10.1207/s15374424jccp3203_15.

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining r-squared measures. *Psychological Methods*, *24*(3), 309–338. doi: 10.1037/met0000184.

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. H. Spieler & E. Schumacher (Eds.), *New methods in cognitive psychology* (pp. 4–31). Psychology Press.

Smith, C. E., & Cribbie, R. (2014). Factorial anova with unbalanced data: A fresh look at the types of sums of squares. *Journal of Data Science*, *12*, 385–404. doi: 10.6339/JDS.201407_12(3).0001.

Streiner, D. L. (2019). Commentary no. 31: The uses and misuses of the analysis of covariance. *Journal of Clinical Psychopharmacology*, *39*(2), 97–99. doi: 10.1097/jcp.0000000000001009.

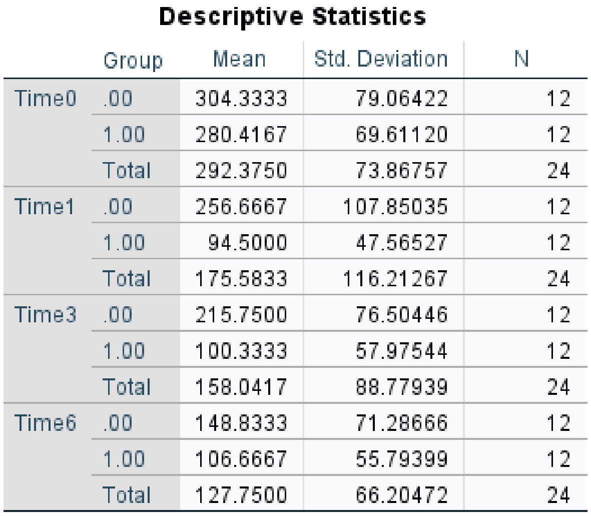
UCLA Statistical Consulting Group. (n.d.). *Repeated measures analysis with SPSS*. https://stats.idre.ucla.edu/spss/seminars/repeated-measures/

Vickers, A. J. (2005). Analysis of variance is easily misapplied in the analysis of randomized trials: A critique and discussion of statistical alternatives. *Psychosomatic Medicine*, *67*(4), 652–655. doi: 10.1097/01.psy.0000172624.52957.a8.

White, I. R., Horton, N. J., Carpenter, J., & Pocock, S. J. (2011). Strategy for intention to treat analysis in randomised trials with missing outcome data. *BMJ*, *342*(d40), 1–5. doi: 10.1136/bmj.d40.

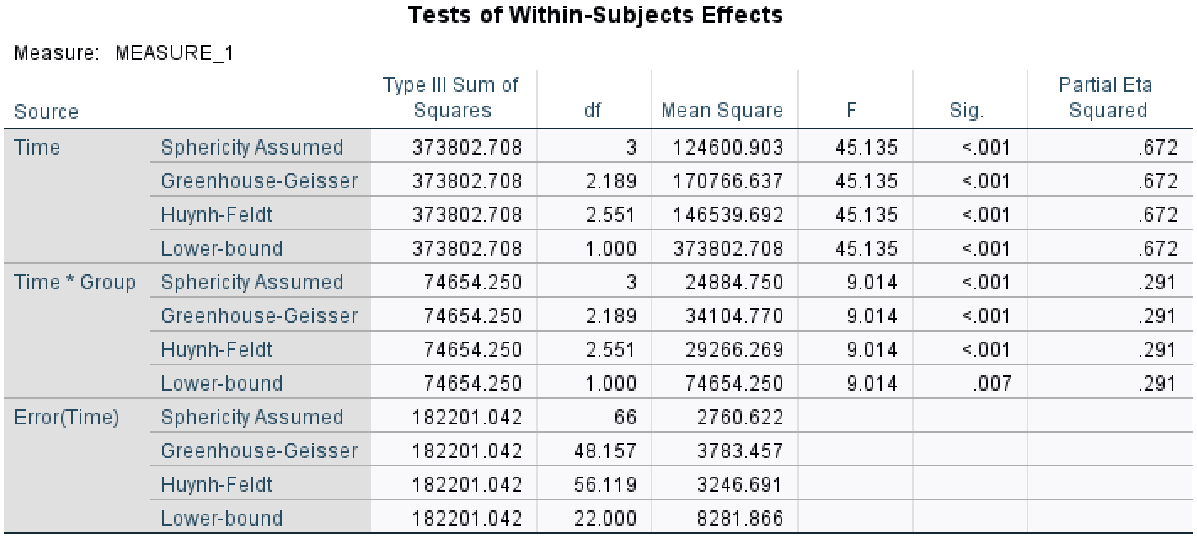
Appendix: The Outputs

Output 1 ■ Mixed Model ANOVA Output



Descriptive Statistics

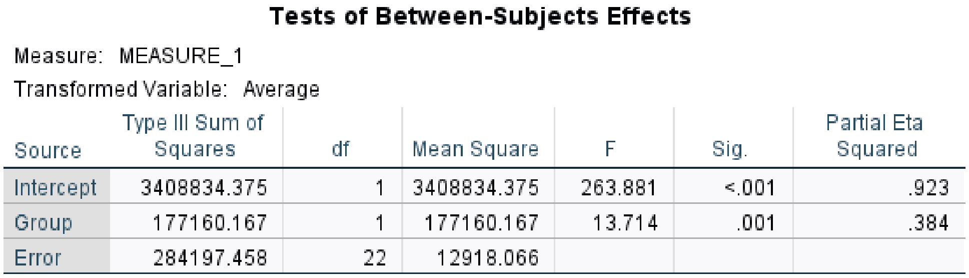
	Group	Mean	Std. Deviation	N
Time0	.00	304.3333	79.06422	12
	1.00	280.4167	69.61120	12
	Total	292.3750	73.86757	24
Time1	.00	256.6667	107.85035	12
	1.00	94.5000	47.56527	12
	Total	175.5833	116.21267	24
Time3	.00	215.7500	76.50446	12
	1.00	100.3333	57.97544	12
	Total	158.0417	88.77939	24
Time6	.00	148.8333	71.28666	12
	1.00	106.6667	55.79399	12
	Total	127.7500	66.20472	24



Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Time	Sphericity Assumed	373802.708	3	124600.903	45.135	<.001	.672
	Greenhouse-Geisser	373802.708	2.189	170766.637	45.135	<.001	.672
	Huynh-Feldt	373802.708	2.551	146539.692	45.135	<.001	.672
	Lower-bound	373802.708	1.000	373802.708	45.135	<.001	.672
Time * Group	Sphericity Assumed	74654.250	3	24884.750	9.014	<.001	.291
	Greenhouse-Geisser	74654.250	2.189	34104.770	9.014	<.001	.291
	Huynh-Feldt	74654.250	2.551	29266.269	9.014	<.001	.291
	Lower-bound	74654.250	1.000	74654.250	9.014	.007	.291
Error(Time)	Sphericity Assumed	182201.042	66	2760.622			
	Greenhouse-Geisser	182201.042	48.157	3783.457			
	Huynh-Feldt	182201.042	56.119	3246.691			
	Lower-bound	182201.042	22.000	8281.866			



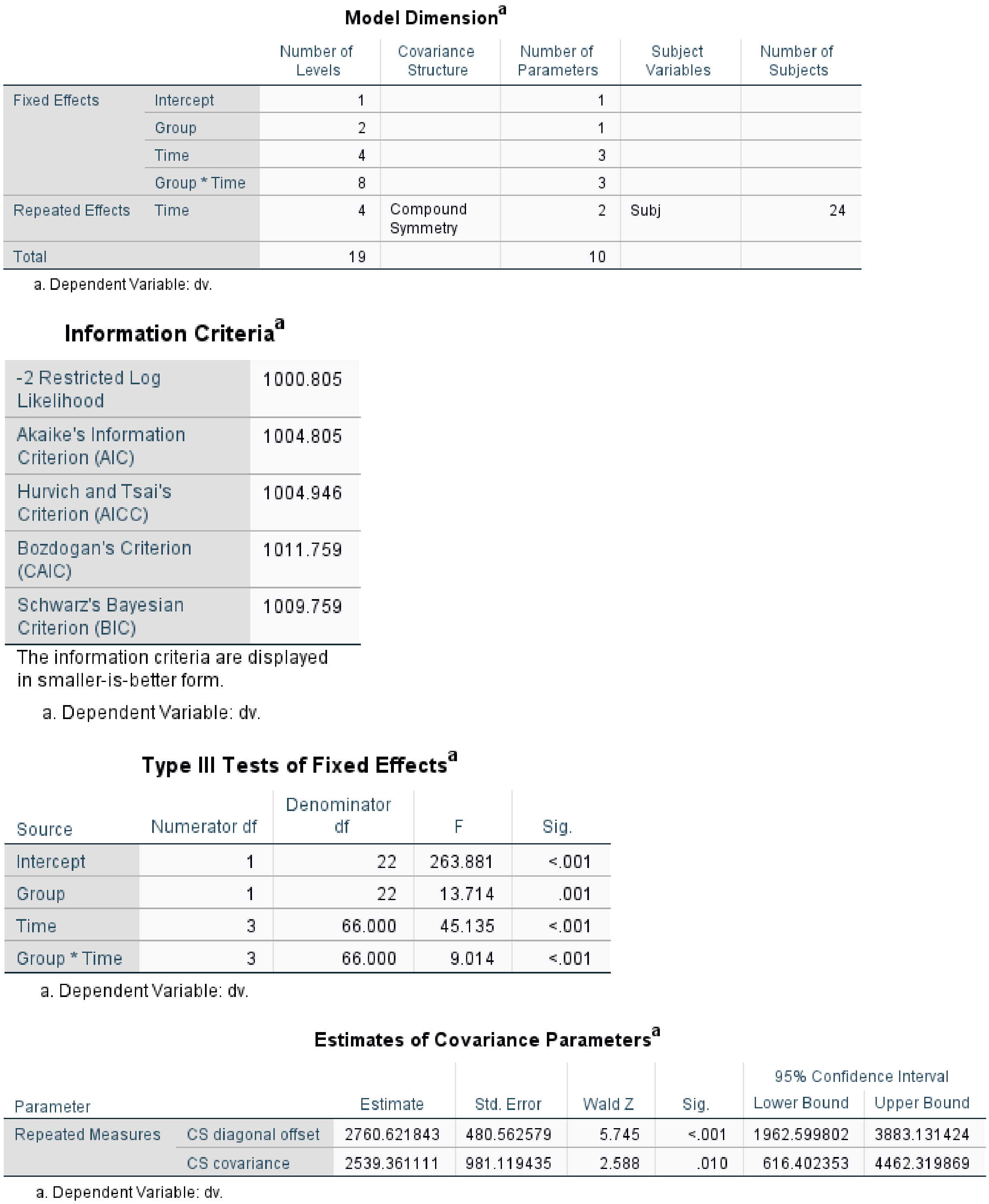
Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	3408834.375	1	3408834.375	263.881	<.001	.923
Group	177160.167	1	177160.167	13.714	.001	.384
Error	284197.458	22	12918.066			

Output 2 ■ *Marginal Model Results for the Repeated Model (CS)*



Model Dimension[a]

		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables	Number of Subjects
Fixed Effects	Intercept	1		1		
	Group	2		1		
	Time	4		3		
	Group * Time	8		3		
Repeated Effects	Time	4	Compound Symmetry	2	Subj	24
Total		19		10		

a. Dependent Variable: dv.

Information Criteria[a]

-2 Restricted Log Likelihood	1000.805
Akaike's Information Criterion (AIC)	1004.805
Hurvich and Tsai's Criterion (AICC)	1004.946
Bozdogan's Criterion (CAIC)	1011.759
Schwarz's Bayesian Criterion (BIC)	1009.759

The information criteria are displayed in smaller-is-better form.

a. Dependent Variable: dv.

Type III Tests of Fixed Effects[a]

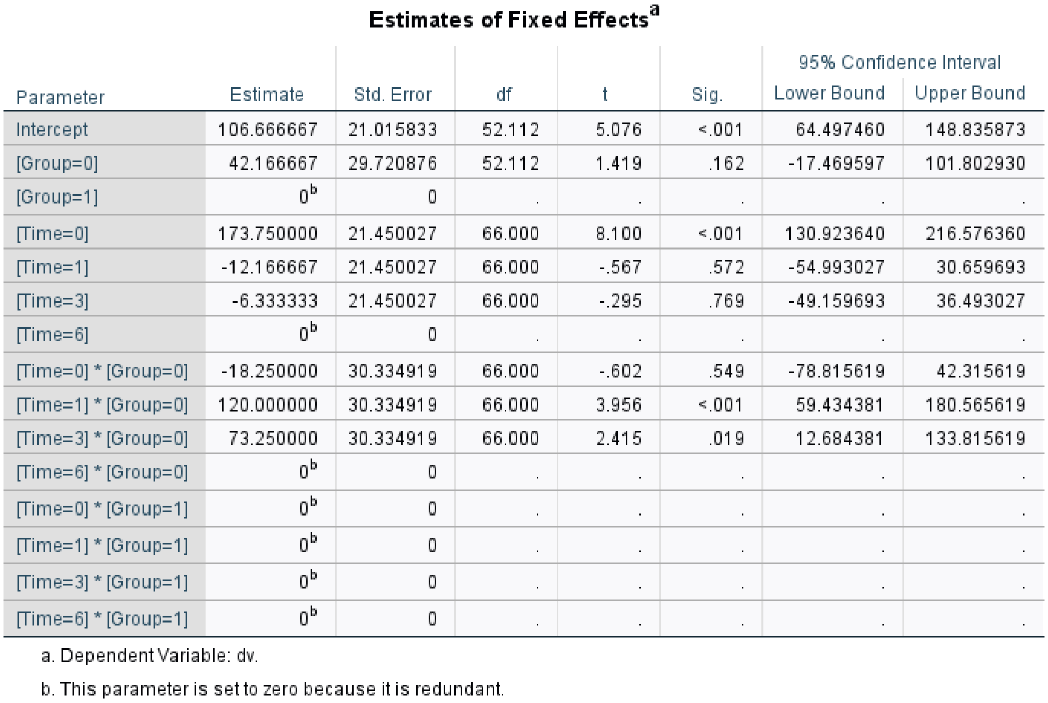
Source	Numerator df	Denominator df	F	Sig.
Intercept	1	22	263.881	<.001
Group	1	22	13.714	.001
Time	3	66.000	45.135	<.001
Group * Time	3	66.000	9.014	<.001

a. Dependent Variable: dv.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Repeated Measures	CS diagonal offset	2760.621843	480.562579	5.745	<.001	1962.599802	3883.131424
	CS covariance	2539.361111	981.119435	2.588	.010	616.402353	4462.319869

a. Dependent Variable: dv.

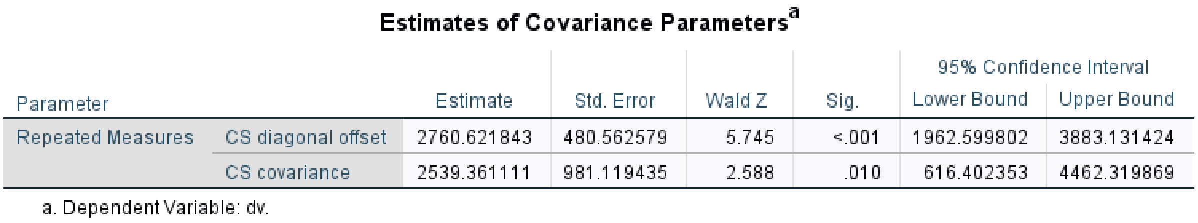


Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	106.666667	21.015833	52.112	5.076	<.001	64.497460	148.835873
[Group=0]	42.166667	29.720876	52.112	1.419	.162	-17.469597	101.802930
[Group=1]	0[b]	0	.	.	.	.	.
[Time=0]	173.750000	21.450027	66.000	8.100	<.001	130.923640	216.576360
[Time=1]	-12.166667	21.450027	66.000	-.567	.572	-54.993027	30.659693
[Time=3]	-6.333333	21.450027	66.000	-.295	.769	-49.159693	36.493027
[Time=6]	0[b]	0	.	.	.	.	.
[Time=0] * [Group=0]	-18.250000	30.334919	66.000	-.602	.549	-78.815619	42.315619
[Time=1] * [Group=0]	120.000000	30.334919	66.000	3.956	<.001	59.434381	180.565619
[Time=3] * [Group=0]	73.250000	30.334919	66.000	2.415	.019	12.684381	133.815619
[Time=6] * [Group=0]	0[b]	0	.	.	.	.	.
[Time=0] * [Group=1]	0[b]	0	.	.	.	.	.
[Time=1] * [Group=1]	0[b]	0	.	.	.	.	.
[Time=3] * [Group=1]	0[b]	0	.	.	.	.	.
[Time=6] * [Group=1]	0[b]	0	.	.	.	.	.

a. Dependent Variable: dv.

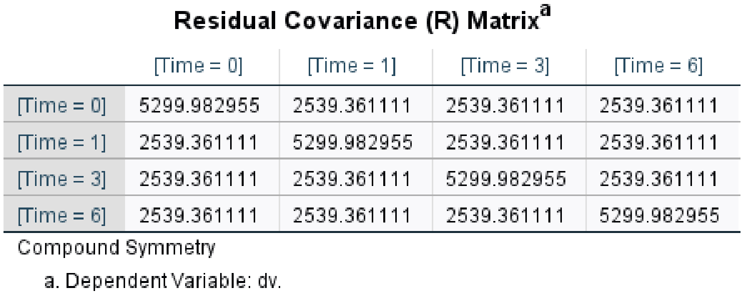
b. This parameter is set to zero because it is redundant.



Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Repeated Measures	CS diagonal offset	2760.621843	480.562579	5.745	<.001	1962.599802	3883.131424
	CS covariance	2539.361111	981.119435	2.588	.010	616.402353	4462.319869

a. Dependent Variable: dv.



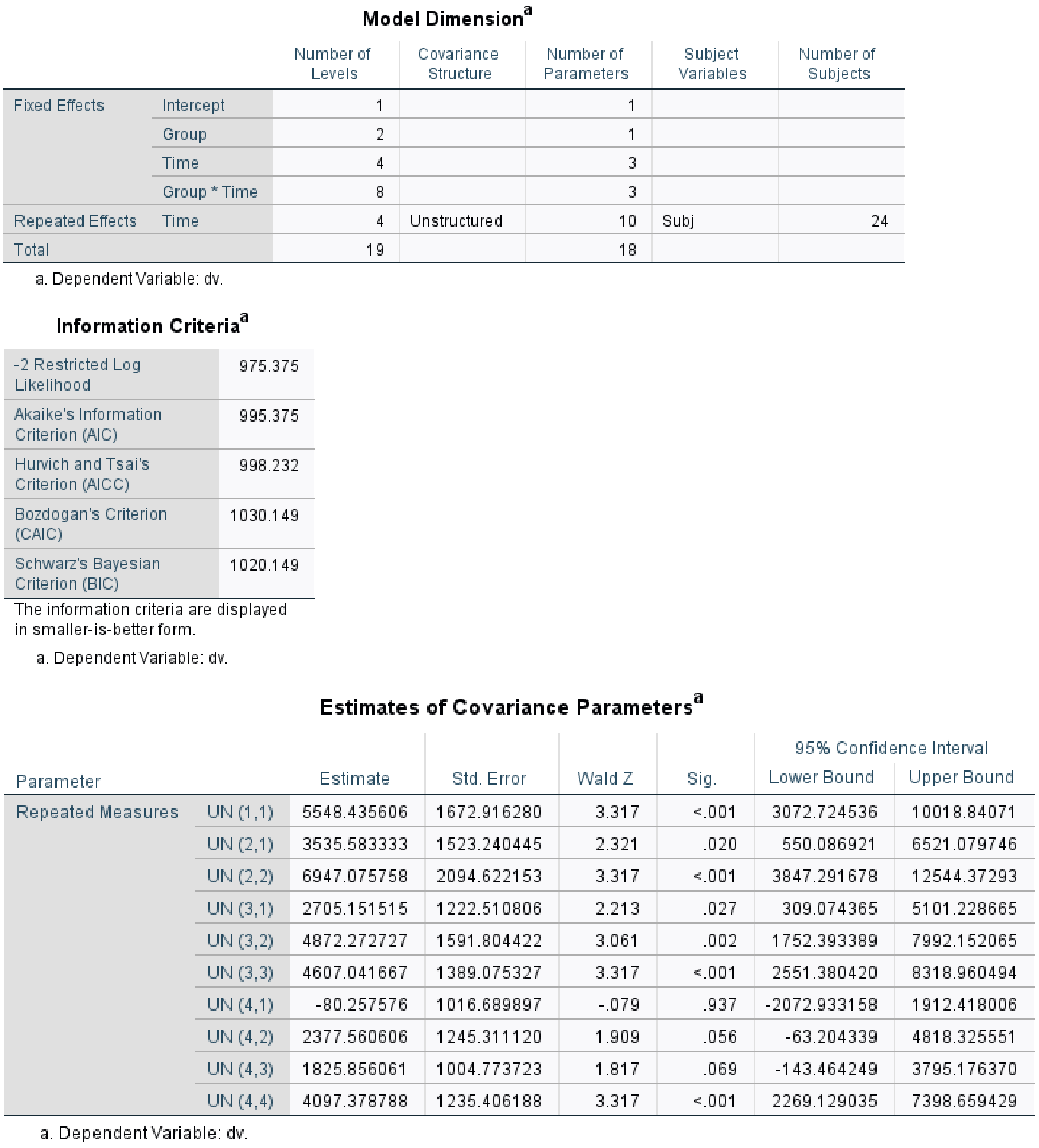
Residual Covariance (R) Matrix[a]

	[Time = 0]	[Time = 1]	[Time = 3]	[Time = 6]
[Time = 0]	5299.982955	2539.361111	2539.361111	2539.361111
[Time = 1]	2539.361111	5299.982955	2539.361111	2539.361111
[Time = 3]	2539.361111	2539.361111	5299.982955	2539.361111
[Time = 6]	2539.361111	2539.361111	2539.361111	5299.982955

Compound Symmetry

a. Dependent Variable: dv.

Output 3 ■ *Selected Marginal Model Results for the Repeated Model (UN)*



Model Dimension[a]

		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables	Number of Subjects
Fixed Effects	Intercept	1		1		
	Group	2		1		
	Time	4		3		
	Group * Time	8		3		
Repeated Effects	Time	4	Unstructured	10	Subj	24
Total		19		18		

a. Dependent Variable: dv.

Information Criteria[a]

-2 Restricted Log Likelihood	975.375
Akaike's Information Criterion (AIC)	995.375
Hurvich and Tsai's Criterion (AICC)	998.232
Bozdogan's Criterion (CAIC)	1030.149
Schwarz's Bayesian Criterion (BIC)	1020.149

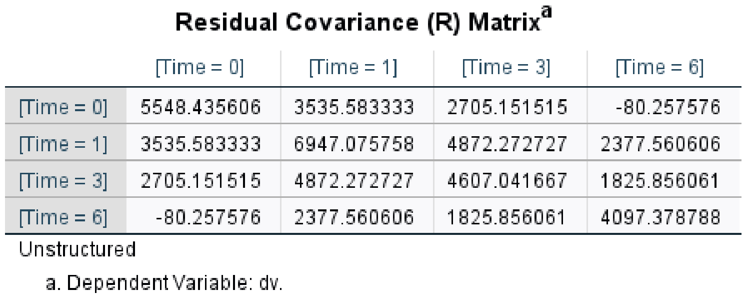
The information criteria are displayed in smaller-is-better form.

a. Dependent Variable: dv.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Repeated Measures	UN (1,1)	5548.435606	1672.916280	3.317	<.001	3072.724536	10018.84071
	UN (2,1)	3535.583333	1523.240445	2.321	.020	550.086921	6521.079746
	UN (2,2)	6947.075758	2094.622153	3.317	<.001	3847.291678	12544.37293
	UN (3,1)	2705.151515	1222.510806	2.213	.027	309.074365	5101.228665
	UN (3,2)	4872.272727	1591.804422	3.061	.002	1752.393389	7992.152065
	UN (3,3)	4607.041667	1389.075327	3.317	<.001	2551.380420	8318.960494
	UN (4,1)	-80.257576	1016.689897	-.079	.937	-2072.933158	1912.418006
	UN (4,2)	2377.560606	1245.311120	1.909	.056	-63.204339	4818.325551
	UN (4,3)	1825.856061	1004.773723	1.817	.069	-143.464249	3795.176370
	UN (4,4)	4097.378788	1235.406188	3.317	<.001	2269.129035	7398.659429

a. Dependent Variable: dv.



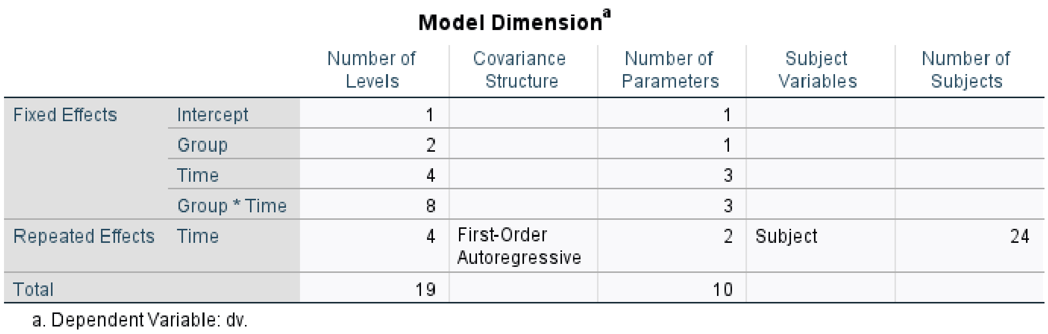
Residual Covariance (R) Matrix[a]

	[Time = 0]	[Time = 1]	[Time = 3]	[Time = 6]
[Time = 0]	5548.435606	3535.583333	2705.151515	-80.257576
[Time = 1]	3535.583333	6947.075758	4872.272727	2377.560606
[Time = 3]	2705.151515	4872.272727	4607.041667	1825.856061
[Time = 6]	-80.257576	2377.560606	1825.856061	4097.378788

Unstructured

a. Dependent Variable: dv.

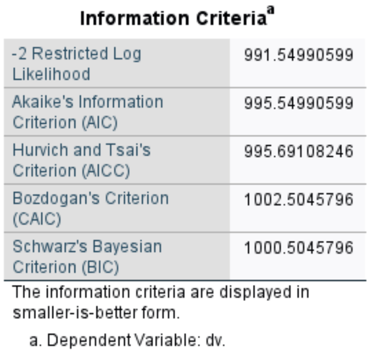
Output 4 ■ Selected Marginal Model Results for the Repeated Model (AR1)



Model Dimension[a]

		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables	Number of Subjects
Fixed Effects	Intercept	1		1		
	Group	2		1		
	Time	4		3		
	Group * Time	8		3		
Repeated Effects	Time	4	First-Order Autoregressive	2	Subject	24
Total		19		10		

a. Dependent Variable: dv.

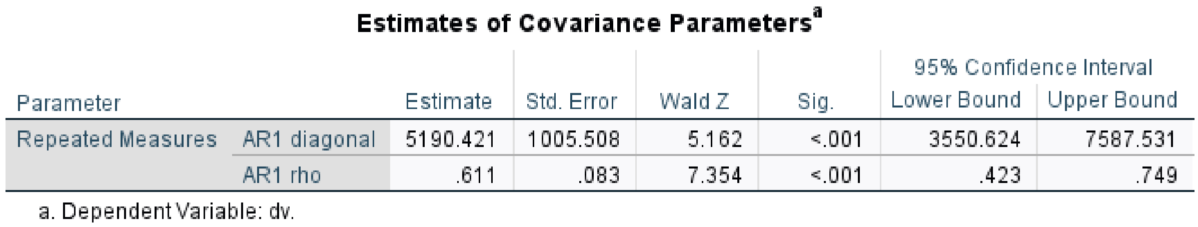


Information Criteria[a]

-2 Restricted Log Likelihood	991.54990599
Akaike's Information Criterion (AIC)	995.54990599
Hurvich and Tsai's Criterion (AICC)	995.69108246
Bozdogan's Criterion (CAIC)	1002.5045796
Schwarz's Bayesian Criterion (BIC)	1000.5045796

The information criteria are displayed in smaller-is-better form.

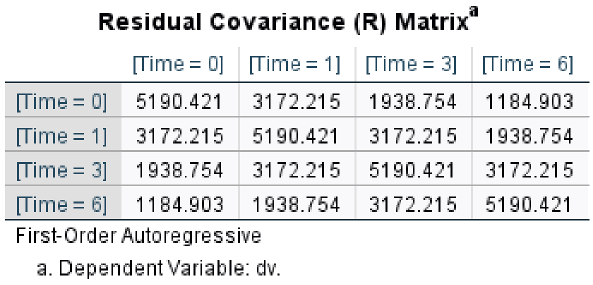
a. Dependent Variable: dv.



Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval Lower Bound	Upper Bound
Repeated Measures	AR1 diagonal	5190.421	1005.508	5.162	<.001	3550.624	7587.531
	AR1 rho	.611	.083	7.354	<.001	.423	.749

a. Dependent Variable: dv.



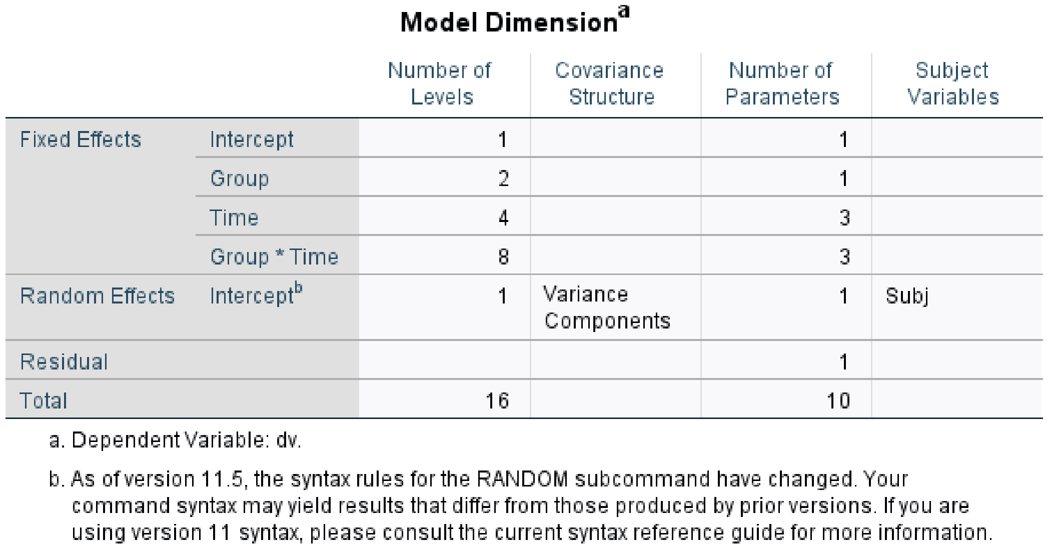
Residual Covariance (R) Matrix[a]

	[Time = 0]	[Time = 1]	[Time = 3]	[Time = 6]
[Time = 0]	5190.421	3172.215	1938.754	1184.903
[Time = 1]	3172.215	5190.421	3172.215	1938.754
[Time = 3]	1938.754	3172.215	5190.421	3172.215
[Time = 6]	1184.903	1938.754	3172.215	5190.421

First-Order Autoregressive

a. Dependent Variable: dv.

Output 5 ■ Random Intercept Model Estimates of Fixed Effects and Covariance Parameters

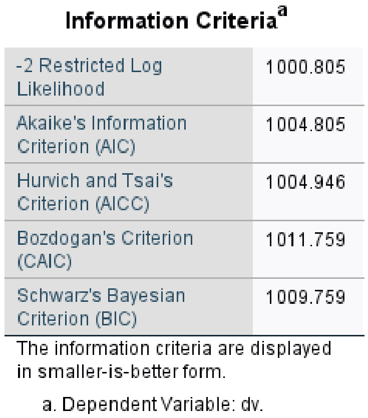


Model Dimension[a]

		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables
Fixed Effects	Intercept	1		1	
	Group	2		1	
	Time	4		3	
	Group * Time	8		3	
Random Effects	Intercept[b]	1	Variance Components	1	Subj
Residual				1	
Total		16		10	

a. Dependent Variable: dv.

b. As of version 11.5, the syntax rules for the RANDOM subcommand have changed. Your command syntax may yield results that differ from those produced by prior versions. If you are using version 11 syntax, please consult the current syntax reference guide for more information.

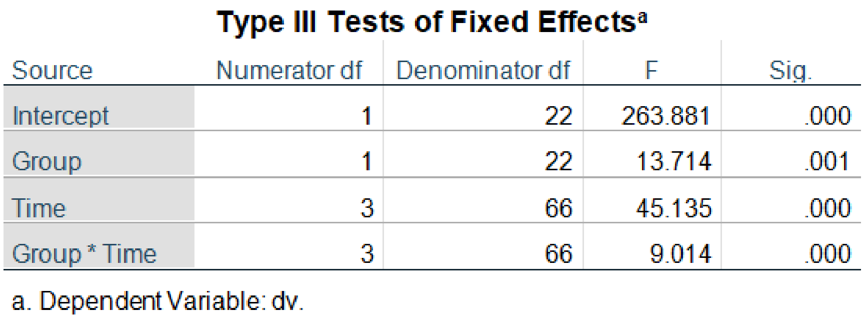


Information Criteria[a]

-2 Restricted Log Likelihood	1000.805
Akaike's Information Criterion (AIC)	1004.805
Hurvich and Tsai's Criterion (AICC)	1004.946
Bozdogan's Criterion (CAIC)	1011.759
Schwarz's Bayesian Criterion (BIC)	1009.759

The information criteria are displayed in smaller-is-better form.

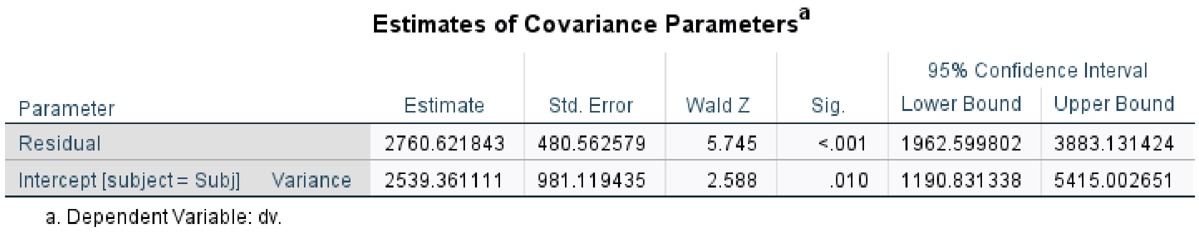
a. Dependent Variable: dv.



Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	22	263.881	.000
Group	1	22	13.714	.001
Time	3	66	45.135	.000
Group * Time	3	66	9.014	.000

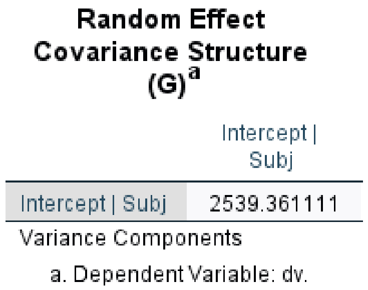
a. Dependent Variable: dv.



Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		2760.621843	480.562579	5.745	<.001	1962.599802	3883.131424
Intercept [subject = Subj]	Variance	2539.361111	981.119435	2.588	.010	1190.831338	5415.002651

a. Dependent Variable: dv.

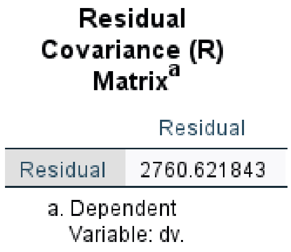


Random Effect Covariance Structure (G)[a]

	Intercept | Subj
Intercept | Subj	2539.361111

Variance Components

a. Dependent Variable: dv.

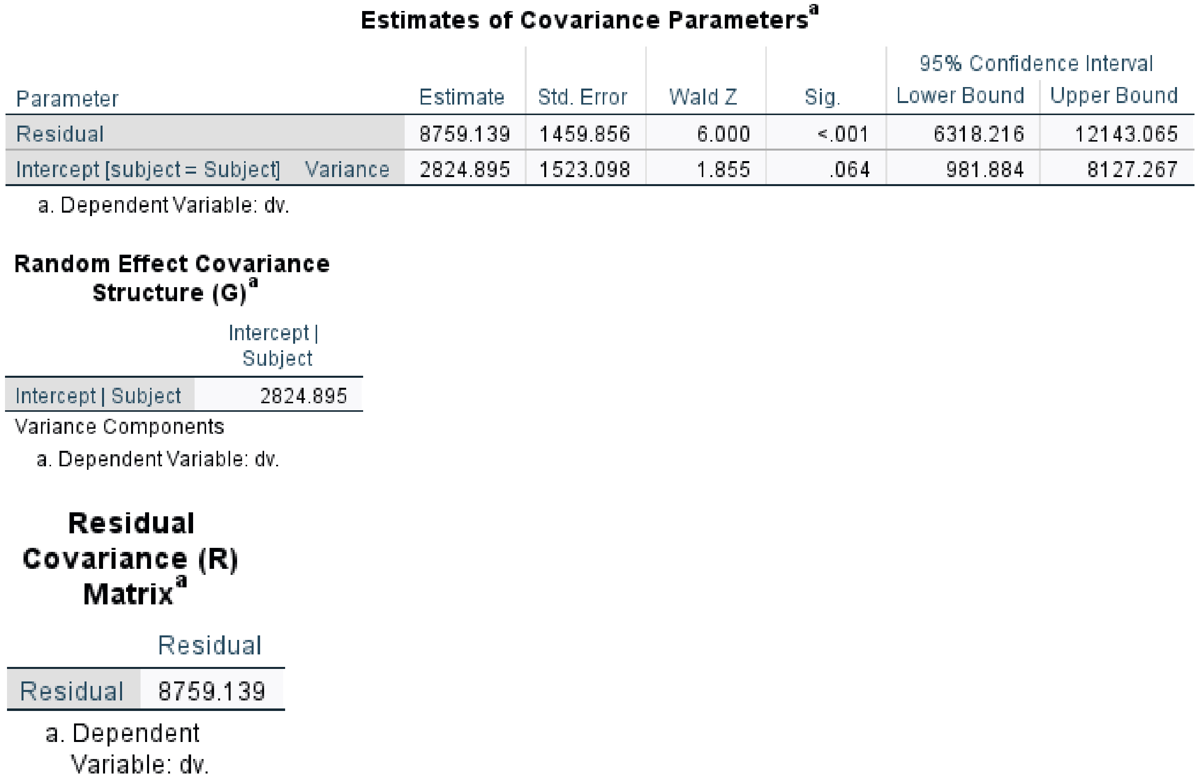


Residual Covariance (R) Matrix[a]

	Residual
Residual	2760.621843

a. Dependent Variable: dv.

Output 6 ■ *Empty Model Estimates of Covariance Parameters*



Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		8759.139	1459.856	6.000	<.001	6318.216	12143.065
Intercept [subject = Subject]	Variance	2824.895	1523.098	1.855	.064	981.884	8127.267

a. Dependent Variable: dv.

Random Effect Covariance Structure (G)[a]

	Intercept | Subject
Intercept | Subject	2824.895

Variance Components

a. Dependent Variable: dv.

Residual Covariance (R) Matrix[a]

	Residual
Residual	8759.139

a. Dependent Variable: dv.

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on osf.io/z5n8a/

Citation

Sharpe, D., & Cribbie, R. A. (2023). Analysis of treatment-control pre-post-follow-up design data. *The Quantitative Methods for Psychology*, *19*(1), 25–46. doi: 10.20982/tqmp.19.1.p025.

Received: 07/02/2022 ∼ Accepted: 29/10/2022