MULTIVARIATE APPLICATIONS SERIES

Longitudinal Analysis

Modeling Within-Person Fluctuation and Change

Lesa Hoffman

ROUTLEDGE

CHAPTER 3

INTRODUCTION TO WITHIN-PERSON ANALYSIS AND MODEL COMPARISONS

Although chapter 1 introduced some of the themes, concepts, and vocabulary recurring in longitudinal analysis, chapter 2 focused exclusively on *between-person models* instead. More specifically, chapter 2 reviewed the general linear models that would be used for cross-sectional data (i.e., in which each person has only one outcome and thus one model residual). We will build on these ideas to continue into chapter 3, whose purpose is to introduce **within-person models**, in which each person has more than one observation of the outcome variable because of repeated measurements (e.g., over time in longitudinal studies; across trials or conditions in other types of repeated measures studies), and thus more than one model residual. Furthermore, although the primary focus of chapter 2 was on the *model for the means* (i.e., interpreting fixed main effects and interactions among predictors), the primary focus of chapter 3 will be on the *model for the variance* (i.e., how the model residuals are distributed and related across observations).

Accordingly, this chapter begins using a two-occasion example to distinguish *between-person* from *within-person* models, and in doing so introduces the *intraclass correlation,* a useful way of quantifying the proportion of between-person variance in a longitudinal outcome. This chapter then describes how the general linear model can be extended for longitudinal analysis via different variants of repeated measures analysis of variance (ANOVA), including the univariate model, the univariate model with adjustments, and the multivariate model. It also discusses the limitations of these models in their suitability for real-world longitudinal data. Finally, this chapter introduces the rules for comparing alternative models, and illustrates these rules by comparing across the variants of ANOVA models seen so far. These rules of model comparisons will be applied throughout the text; a more technical treatment is also provided in chapter 5.

1. Extending Between-Person Models to Within-Person Models

As presented in chapter 1, longitudinal data are unique in their capacity to provide information about between-person (BP) relationships and within-person (WP) relationships simultaneously. This is accomplished by quantifying and predicting the variation in the outcome due to each source, as will be illustrated using simulated two-occasion data in this section.

Consider a hypothetical example: A researcher is interested in examining the effects of a new approach to instruction on an elementary student learning outcome (M = 53.34, SD = 6.35, range = 37.54 to 68.62). She randomly assigns 25 students to a control group (group = 1) and 25 students to a treatment group (group = 2), and collects student data at the beginning of the semester (pre-test is time = 1) and again at the end of the semester (post-test is time = 2), with no missing data. Individual trajectories are shown in Figure 3.1. She hypothesizes that students will score higher on the outcome at post-test than at pre-test (i.e., a positive main effect of time), and also that time and group will have a positive interaction, such that the change over time will be greater for students in the treatment group than in the control group, or equivalently, that the difference between the control and treatment groups will be greater at post-test than at pre-test (and there should be no group difference at pre-test). Let us now examine two different ways we might test these hypotheses in longitudinal data: by using a between-person model for the variance or a within-person model for the variance.

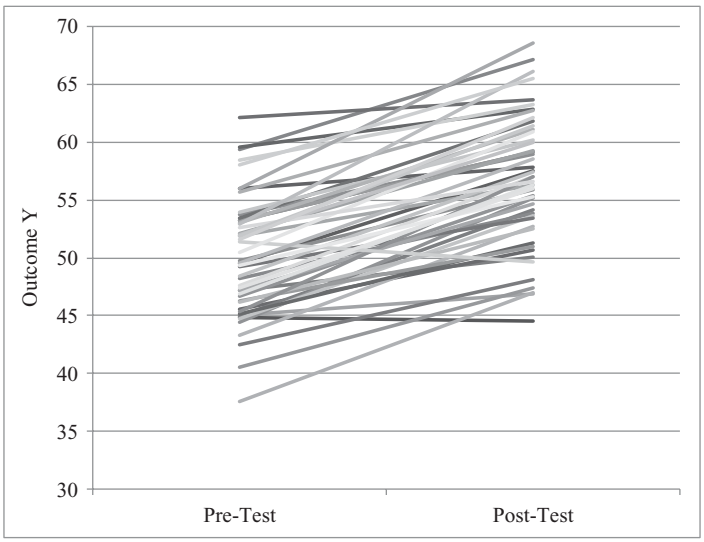

Figure 3.1 Individual trajectories for two-occasion example learning data.

1.A. A Between-Person Empty Model

We can begin by considering the simplest possible model for any outcome variable, a **between-person empty model**, as introduced in chapter 2 and as shown in Equation (3.1):

$$y_{ti} = \beta_0 + e_{ti} \tag{3.1}$$

in which y_{ti} is the outcome at time *t* for individual *i*, β_0 is the intercept, and e_{ti} is the deviation from the intercept at time *t* for individual *i*. In estimating this model for our example data, as shown in the first set of columns of Table 3.1, the intercept $\beta_0 = 53.34$, which is just the grand mean of the outcome over all persons and occasions because we have no predictors included in the model for the means so far. The e_{ti} variance was estimated as $\sigma^2_e = 40.34$, which so far includes all possible variation in the outcome across persons and occasions.

However, to what extent is this between-person model containing only a single residual appropriate for these longitudinal data? Recall the assumptions of the σ^2_e-only between-person models (e.g., between-groups analysis of variance, regression) as introduced in chapter 2: the e_{ti} residuals are supposed to be conditionally normally distributed (i.e., after including model predictors) with constant variance (across all observations and predictors) and to be independent across observations. Thus, the model in Equation (3.1) assumes that participants varied just as much from each other at pre-test as at post-test (i.e., constant variance across time), which may not be likely. More problematically, it also assumes that the residuals for the pre-test and post-test occasions from the same person have no relationship at all. This independence assumption is highly unlikely to hold in longitudinal data, even for just two occasions. Fortunately, this independence assumption is testable. To do so, we need to augment our between-person model for the variance to be able to represent and quantify the separate contributions of cross-sectional, between-person variation as well as longitudinal, within-person variation, as presented next.

1.B. A Within-Person Empty Model

A **within-person model** uses two parameters in the model for the variance to distinguish between-person (BP) from within-person (WP) variance in longitudinal data. Note that the within-person model does not remove between-person variation, but instead includes *both* sources of variation simultaneously. A **within-person empty model** is shown in Equation (3.2):

$$y_{ti} = \beta_0 + U_{0i} + e_{ti} \tag{3.2}$$

in which the model for the means includes just the intercept β_0, which is now the mean of the person means given that the model for the variance now contains two residual terms to represent the variance in the outcome, U_{0i} and e_{ti}. The interpretation of these residual terms is illustrated in Figure 3.2, which contrasts how

Table 3.1 Results for the two-occasion example data from the between-person and within-person models. Bold values are $p < .05$.

Model Parameters		*Equation 3.1: Between-Person Empty Model*			*Equation 3.2: Within-Person Empty Model*			*Equation 3.6: Between-Person Conditional Model*			*Equation 3.7: Within-Person Conditional Model*		
		Est	*SE*	$p <$	*Est*	*SE*	$p <$	*Est*	*SE*	$p <$	*Est*	*SE*	$p <$
Model for the Means													
β_0	Intercept	**53.34**	0.64	.001	**53.34**	0.73	.001	**49.08**	1.04	.001	**49.08**	1.04	.001
	Time Effect												
β_1	Control Group							**5.82**	1.48	.001	**5.82**	0.60	.001
$\beta_1 + \beta_3$	Treatment Group							**7.86**	1.48	.001	**7.86**	0.60	.001
	Group Effect												
β_2	Pre-Test							1.68	1.48	.260	1.68	1.48	.260
$\beta_2 + \beta_3$	Post-Test							**3.72**	1.48	.015	**3.72**	1.48	.015
	Time by Group Interaction												
β_3	Difference of Difference							2.04	2.09	.333	**2.04**	0.84	.019
Predicted Cell Means													
β_0	= Control Group, Pre-Test							**49.08**	1.04	.001	**49.08**	1.04	.001
$\beta_0 + \beta_1$	= Control Group, Post-Test							**54.90**	1.04	.001	**54.90**	1.04	.001
$\beta_0 + \beta_2$	= Treatment Group, Pre-Test							**50.76**	1.04	.001	**50.76**	1.04	.001
$\beta_0 + \beta_1 + \beta_2 + \beta_3$	= Treatment Group, Post-Test							**58.62**	1.04	.001	**58.62**	1.04	.001
Model for the Variance													
σ^2_e	Residual Variance	**40.34**	5.73	.001	**28.21**	5.64	.001	**27.22**	3.93	.001	**4.45**	0.91	.001
$\tau^2_{U_0}$	Random Intercept Variance				**12.25**	6.03	.042				**22.78**	5.12	.001
	Intraclass Correlation	.00			.30			.00			.84		

the residuals are defined in a *between-person empty model* (top panel) versus a *within-person empty model* (bottom panel). In both panels, the dashed line through the squares represents prediction of the grand mean outcome from the intercept β_0. Outcomes for two hypothetical persons (A and B) are shown by the dots and triangles, respectively.

Let us first consider the outcomes for Person A in the top panel of Figure 3.2, in which the differences between the actual outcomes and the outcomes predicted from the model for the means (just β_0 so far) are represented by the e_{ti} deviations. Because Person A performs better than average at both pre-test and post-test, Person A's e_{ti} residuals are likely to be correlated (or *dependent*) as a result. The same is true for the outcomes for Person B, whose outcomes are both below average, and

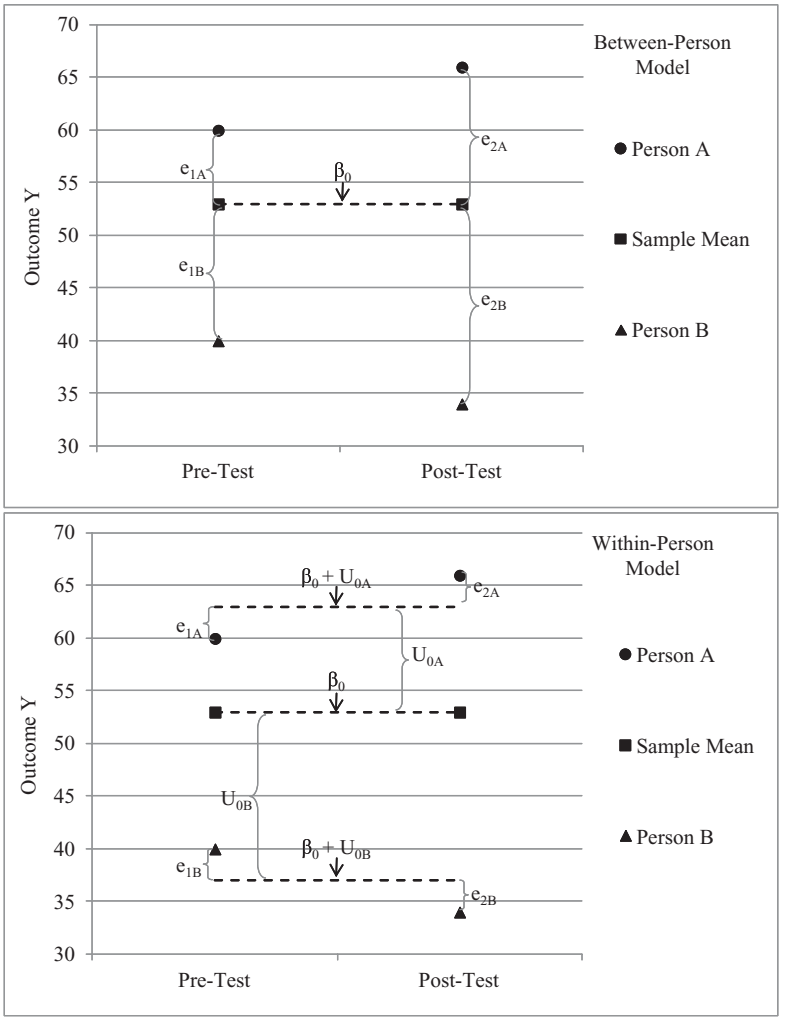

Figure 3.2 Illustration of empty between-person (top) and within-person (bottom) models for the variance. The data for the sample and for two hypothetical individuals are shown.

whose e_{ti} residuals will also be dependent. As you might remember from chapter 1, this illustrates one of several kinds of dependency possible in longitudinal data—correlation among the residuals from the same person due to constant mean differences over time. The trick to solving this problem is to add a residual term to the model for the variance that will accurately represent this type of dependency, rather than assuming no dependency at all.

Accordingly, this dependency due to constant mean differences can be included in the model for the variance as U_{0i} as shown in the bottom panel of Figure 3.2. This new term, U_{0i}, is called a **random intercept**, and is *the difference between the conditional mean predicted by the model for the means (here, just β_0) and the person's mean across time*. As a result, we must now explicitly differentiate two kinds of intercepts: not only do we have the *fixed* intercept β_0 that is shared by all observations in the sample, but now we also have a *random* intercept U_{0i}, by which each person gets his or her own *deviation* from the fixed intercept. The β_0 and U_{0i} terms share a *0* subscript because they are both intercepts, but only the random intercept U_{0i} gets an *i* subscript to indicate that each person has a different U_{0i}. In the bottom panel of Figure 3.2, $\beta_0 = 53$ (still the grand mean over time). For Person A, $U_{0i} = 10$, such that $\beta_0 + U_{0A} = 53 + 10 = 63$, Person A's mean across time. Similarly, for Person B, $U_{0i} = -16$, so $\beta_0 + U_{0B} = 53 - 16 = 37$, Person B's mean across time. Although the individual U_{0i} values could be calculated, longitudinal models instead estimate the variance of the U_{0i} values across persons as $\tau^2_{U_0}$. That is, because U_{0i} is considered a random variable (i.e., just like e_{ti} is a random variable), it doesn't really matter what the individual U_{0i} values are. As such, although the residual term U_{0i} is included in the model equations to represent the *idea* of between-person variance in the mean over time, its variance $\tau^2_{U_0}$ is the actual model parameter to be estimated. In keeping with the notation used in other books (e.g., Raudenbush & Bryk, 2002; Snijders & Bosker, 2012), τ^2_U will be used for variances of random effects (i.e., just U_{0i} for now), and σ^2_e will be used for the variance of the e_{ti} residuals.

After accounting for dependency due to between-person mean differences by including the random intercept U_{0i}, the remaining within-person deviation of the outcome at each occasion from the person's mean is then represented by e_{ti}. Thus, for Person A, e_{ti} at pre-test would be –3, such that the pre-test outcome is created by $\beta_0 + U_{0A} + e_{1A} = 53 + 10 - 3 = 60$, whereas e_{ti} at post-test would be 3, such that the post-test outcome is created by $\beta_0 + U_{0A} + e_{2A} = 53 + 10 + 3 = 66$. For Person B, the pre-test outcome is created by $\beta_0 + U_{0B} + e_{1B} = 53 - 16 + 3 = 40$, and the post-test outcome is created by $\beta_0 + U_{0B} + e_{2B} = 53 - 16 - 3 = 34$. Both the U_{0i} and e_{ti} residuals are assumed to be conditionally normally distributed with constant variance across persons and occasions—although we will see how the latter assumption about constant variance across occasions can be relaxed in the models presented in the next few chapters.

It is important to note that the total amount of variation around the grand mean at each occasion will be approximately the same when including only e_{ti} (as in the between-person empty model in the top panel of Figure 3.2) as when including U_{0i} and e_{ti} (as in the within-person empty model in the bottom panel of Figure 3.2). Because the within-person empty model in Equation (3.2) contains only the fixed intercept β_0 in the model for the means (thus the name "empty"), we have

not yet *explained* any outcome variance. Instead, we have simply *partitioned* the total outcome variance into different sources for its two dimensions of sampling: variation between persons in $\tau^2_{U_0}$ and variation within persons in σ^2_e, rather than keeping all of the outcome variance in σ^2_e. In doing so we have allowed a correlation across occasions, as explained next.

1.C. Intraclass Correlation

As shown in the second set of columns of Table 3.1, estimating the within-person empty model in Equation (3.2) in our example yields a random intercept variance of the U_{0i} values of $\tau^2_{U_0}$ = 12.25 and a residual variance for the e_{ti} values of σ^2_e = 28.21. Their total variance of 40.46 is slightly higher than the total variance from the between-person empty model of σ^2_e = 40.34. This illustrates how the partitioning of variance in the within-person empty model (into $\tau^2_{U_0}$ and σ^2_e) will result in a total variance that is approximately (but not always exactly) the same as the total variance in the between-person empty model (from σ^2_e).

But why should we care what the variances of the U_{0i} and e_{ti} terms are? Although not helpful for the prediction of our outcome, the results of a within-person empty model will provide a baseline with which to judge the contribution of other effects to be added to the model. It also provides a means with which to calculate a useful descriptive statistic, the **intraclass correlation**, or **ICC**, as shown for our example data in Equation (3.3):

$$\begin{aligned} \text{ICC} &= \frac{\text{BP variation}}{\text{BP + WP variation}} = \frac{\text{Var}(U_{0i})}{\text{Var}(U_{0i}) + \text{Var}(e_{ti})} = \frac{\tau^2_{U_0}}{\tau^2_{U_0} + \sigma^2_e} \\ &= \frac{12.25}{12.25 + 28.21} = .30 \end{aligned} \tag{3.3}$$

in which the ICC can range between 0 and 1. One way to interpret the ICC is as *the proportion of outcome variation due to between-person differences in the intercept* (i.e., due to variance in the between-person U_{0i} random intercepts, relative to the total variation from the between-person U_{0i} random intercepts and the within-person e_{ti} residuals). Thus, 30% of the original outcome variation (i.e., before adding any predictors) is due to between-person mean differences over time. In other words, 30% of the outcome variance is cross-sectional and 70% is longitudinal.

To understand the implications of this ICC value for our data analysis, it may be helpful to consider the results from fitting the between-person (BP) and within-person (WP) empty models to our example data. As shown in Table 3.1, both models result in the exact same fixed intercept of β_0 = 53.34, which is the grand mean over time given that we do not yet have any predictors. This is because both models have specified the same *model for the means*—an empty model, which (so far) excludes any predictors for change over time or mean differences across groups. Instead, the BP and WP models differ in their *model for the variance,* such that the BP model includes σ^2_e only, whereas the WP model includes both $\tau^2_{U_0}$ and σ^2_e. The

implications of this difference with respect to what these models predict are shown in Equation (3.4):

$$\begin{aligned}\text{BP model: } &\begin{pmatrix}\text{Var}(y_1) & \text{Cov}(y_1,y_2)\\ \text{Cov}(y_1,y_2) & \text{Var}(y_2)\end{pmatrix} = \begin{pmatrix}\sigma^2_e & 0\\ 0 & \sigma^2_e\end{pmatrix} = \begin{pmatrix}40.34 & 0\\ 0 & 40.34\end{pmatrix}\\ \text{WP model: } &\begin{pmatrix}\text{Var}(y_1) & \text{Cov}(y_1,y_2)\\ \text{Cov}(y_1,y_2) & \text{Var}(y_2)\end{pmatrix} = \begin{pmatrix}\sigma^2_e + \tau^2_{U_0} & \tau^2_{U_0}\\ \tau^2_{U_0} & \sigma^2_e + \tau^2_{U_0}\end{pmatrix} = \begin{pmatrix}40.46 & 12.25\\ 12.25 & 40.46\end{pmatrix}\end{aligned} \quad (3.4)$$

in which the variance across persons at each occasion (in which y_1 = pre-test and y_2 = post-test) and the covariance between occasions are abbreviated as *Var* and *Cov,* respectively. As a brief refresher, Equation (3.5) provides the formulas for the variance of the outcome at each occasion (left equation) and for the covariance between two occasions (right equation):

$$\text{Variance } (y_t) = \frac{\sum_{i=1}^{N}(y_{ti} - \hat{y}_{ti})^2}{N - k} \quad \text{Covariance } (y_1, y_2) = \frac{\sum_{i=1}^{N}(y_{1i} - \hat{y}_{1i})(y_{2i} - \hat{y}_{2i})}{N - k} \quad (3.5)$$

in which the variance of the outcome at time *t* can be calculated as the sum over *N* persons of the squared deviations of each actual y_{ti} from the $\hat{y}_{ti}$ predicted from the model for the means divided by *N* persons minus *k* fixed effects (*k* = 1 for just β_0 so far), and the covariance between occasions can be calculated as the sum over *N* persons of the product of those $y_{ti} - \hat{y}_{ti}$ deviations at each occasion divided by *N* persons minus *k* fixed effects. Covariance tells us the direction and extent to which the residuals of the pre-test and post-test occasions are related in the unstandardized metric of the outcome. Although these formulas in Equation (3.5) are based on least squares estimation, they are also equivalent in this case to the restricted maximum likelihood results that we will utilize later (because in this example we have complete data that is balanced over time).

For the predicted variance and covariance using the BP model in the top of Equation (3.4), all the variance in the outcome at each occasion is represented by the single residual e_{ti} (which is the difference between y_{ti} and $\hat{y}_{ti}$ for each observation, as shown in the top panel of Figure 3.2). In our example, the BP model predicts the variance at each occasion to be 40.34. It assumes the e_{ti} residuals have a normal distribution and are independent with constant variance across observations. Thus, the BP model predicts no covariance whatsoever between the pre-test and post-test residuals for each person (indicated by 0 for the covariance in the off-diagonal).

In contrast, for the predicted variance and covariance using the WP model shown in the bottom of Equation (3.4), the total outcome variance at each occasion (estimated as 40.46) is predicted from the variance of the e_{ti} residuals (WP variance of $\sigma^2_e = 28.21$) plus the variance of the U_{0i} random intercepts (BP variance of $\tau^2_{U_0} = 12.25$). Like the BP model, the WP model predicts the same total outcome variation at pre-test and post-test. But unlike the BP model, the WP model also predicts a covariance between the pre-test and post-test outcomes, and this covariance is entirely due to the random intercept variance ($\tau^2_{U_0} = 12.25$). Although it may seem odd that a variance could become a covariance, in this case, what the WP model is saying is that *the only reason why the pre-test and post-test outcomes are related is because of constant*

mean differences between persons over time. After including the U_{0i} random intercept to represent those constant person mean differences (i.e., deviating each y_{ti} from $\hat{y}_{ti} + U_{0i}$ in the bottom panel of Figure 3.2 instead of just $\hat{y}_{ti}$ as in the top panel), the e_{ti} residuals from the same person are then independent. Both the BP and WP models assume the residuals from different people are independent as well—although in chapter 11 we'll see how the model for the variance can be further modified if persons are not actually independent, such as when nested in groups.

We can convert the *covariance* between occasions to a *correlation* using Equation (3.6):

$$\begin{aligned}\text{Correlation}\,(y_1, y_2) &= \frac{\text{Cov}(y_1, y_2)}{\sqrt{\text{Var}(y_1)} * \sqrt{\text{Var}(y_2)}} = \frac{\tau^2_{U_0}}{\sqrt{\tau^2_{U_0} + \sigma^2_e} * \sqrt{\tau^2_{U_0} + \sigma^2_e}} \\ &= \frac{12.25}{40.46} = .30\end{aligned} \tag{3.6}$$

resulting in a correlation between occasions of $r = .30$, as reported in the second set of columns of Table 3.1. It is not a coincidence that .30 is what we found for the **intraclass correlation** (ICC) earlier—another way of interpreting the ICC is as *the correlation of the outcome residuals over time*. Although with only two occasions there is only one correlation to consider, once we have more than two occasions, the ICC will become the average correlation across time.

In summary, in our empty models so far, the within-person model with $\tau^2_{U_0}$ and σ^2_e predicts a constant correlation between occasions due to the random intercept U_{0i} (i.e., the only dependency is due to constant person mean differences over time), whereas the between-person model with σ^2_e only predicts no correlation between occasions at all. As a result, a between-person model is not likely to be plausible for longitudinal data, in which the residuals from the same person will usually be dependent because of constant person mean differences, as well as for other reasons, as introduced in chapter 1. But why should we care? That is, the hypotheses in this example concern the effects in the model for the means for time, group, and their interaction, and not how much between-person or within-person variance our outcome has. As we will see next, though, because the model for the variance can result in different inferences about the model's fixed effects, the model for the variance can be just as important to consider.

1.D. Comparing Between-Person and Within-Person Conditional Model Results

Let us now examine the hypothesized effects in the model for the means under between-person (BP) and within-person (WP) versions of the same ***conditional*** model with predictors of time, group, and their interaction, as shown in Equation (3.7):

$$\begin{aligned}\text{BP model: } y_{ti} &= \beta_0 + \beta_1(\text{Time}_{ti}) + \beta_2(\text{Group}_i) + \beta_3(\text{Time}_{ti})(\text{Group}_i) + e_{ti} \\ \text{WP model: } y_{ti} &= \beta_0 + \beta_1(\text{Time}_{ti}) + \beta_2(\text{Group}_i) + \beta_3(\text{Time}_{ti})(\text{Group}_i) \\ &\quad + U_{0i} + e_{ti}\end{aligned} \tag{3.7}$$

in which the BP and WP models differ only in their model for the variance (σ^2_e-only versus $\tau^2_{U_0}$ and σ^2_e, respectively). Their fixed effects in the model for the means are the same: β_0 is the fixed intercept, β_1 is the simple main effect of time, β_2 is the simple main effect of group, and β_3 is the two-way interaction of time by group. The subscript *ti* is used for any variable that varies over both time and individuals (here, the y_{ti} outcome, the predictor of $Time_{ti}$, and the e_{ti} residual). The subscript *i* is used for $Group_i$ because each person is only in one group across all occasions (so group membership varies across individuals, but does not vary over time), as well as for the random intercept U_{0i}, for which each person gets his or her own value (that is constant across time).

As we learned in chapter 2, the interpretation of these fixed effects of time and group and their tests of significance will depend on the reference value for each predictor. In the example dataset, the time variable has possible values of 1 = pre-test and 2 = post-test, and the group variable has values of 1 = control and 2 = treatment. One option is to make the reference point for the model the pre-test observation for the control group by coding the time variable as pre-test = 0 and post-test = 1, and by coding the group variable as control = 0 and treatment = 1. Alternatively, we could make the reference point the treatment group at post-test by coding the time variable as pre-test = 1 and post-test = 0, and by coding the group variable as control = 1 and treatment = 0. There is no wrong way to code the predictor variables, but the choices we make for how they are included will impact the intercept and main effect parameters in model solution we receive (although it will not change the amount of outcome variance the model accounts for, nor will it change the predicted outcome for any observation).

The model results for the BP and WP models in Equation (3.7) are shown in the third and fourth sets of columns of Table 3.1, respectively, along with the model-predicted means for each combination of time and group. The predictors of time and group have been coded so that the reference is the control group at pre-test; thus the fixed intercept $\beta_0 = 49.08$ is predicted outcome for the control group at pre-test. The *simple main effect of time* $\beta_1 = 5.82$ is the difference between the pre-test and the post-test occasions specifically for the control group, which was significant (in both models) as expected. The *simple main effect of group* $\beta_2 = 1.68$ is the difference between the control and treatment groups specifically at pre-test, which was not significant (in either model) as expected. Finally, the two-way time by group interaction $\beta_3 = 2.04$ is the *difference of the differences* and can be interpreted in two alternative but equally correct ways. First, $\beta_3 = 2.04$ is how the difference between the pre-test and post-test occasions differs between the control and treatment groups, such that the change over time is larger by 2.04 in the treatment group. Second, $\beta_3 = 2.04$ is also how the difference between the control and treatment groups differs between pre-test and post-test, such that the group difference is larger by 2.04 at post-test than at pre-test. Interestingly, while the time by group interaction β_3 was not significant in the BP model, it was significant in the WP model. We'll revisit this result shortly.

To examine the significance of the other simple effects not provided directly by the model, we could re-code the time and group predictor variables so that the reference is the treatment group at post-test, or we could use additional programming statements instead (as shown in the example syntax online). In doing so, also as

provided in Table 3.1, we find that the simple main effect of time for the treatment group of $\beta_1 + \beta_3 = 5.82 + 2.04 = 7.86$ and the simple main effect of group at post-test of $\beta_2 + \beta_3 = 1.68 + 2.04 = 3.72$ were also significant (in both models) as expected. Thus, the researcher's hypotheses were partially supported—both the control and treatment groups did improve significantly from pre-test to post-test, and the group difference was not significant at pre-test (i.e., before the treatment) but was significant at post-test (i.e., after the treatment). But the other hypothesis for the time by group interaction β_3—that the treatment group would improve more over time than the control group—was supported in the WP model, but not in the BP model. Why did this happen, and which answer is the right one?

Before answering these questions, we must consider the kind of outcome variance each fixed effect could explain. First, the predictor for control versus treatment group varies between persons (but not within persons because each person is in the same group at both occasions). Thus, including the effect of group in the model should reduce the model's BP variance (the U_{0i} random intercept variance $\tau^2_{U_0}$). Second, the predictor for pre-test versus post-test time varies within persons but not between persons (because each person has the same two possible occasions). Thus, including the effect of time in the model should reduce the model's WP variance (the e_{ti} residual variance σ^2_e). Third, although the interaction of time by group varies both between persons and within persons, it serves to predict why some people change differently from pre-test to post-test. Thus, including the interaction of time by group should further reduce the model's WP variance—because the interaction allows each group to have their own slope for time, the remaining time- and individual-specific e_{ti} deviations around the time slope should be reduced.

With these differences as to whether the predictor is accounting for BP variance or WP variance in mind, let us now note some important similarities and differences in the results in Table 3.1. First, we note that the parameter estimates for the fixed effects (β_0, β_1, β_2, and β_3, and the model-implied additional simple main effects of $\beta_1 + \beta_3$ and $\beta_2 + \beta_3$) are exactly the same across the BP and WP models. This is not a coincidence—the BP and WP models in Equation (3.7) include the same model for the means, and so their estimates of these fixed effects are the same. However, the BP and WP models include different models for their variance. So, what does differ between models are the *standard errors* (and resulting *p*-values) of the effects that account for within-person variance: those for time and time by group. Standard errors (SE) for the main effects or interactions from our models can be calculated as shown in Equation (3.8):

$$\text{BP effect } SE_{\beta_x} = \sqrt{\frac{\sigma^2_e + \tau^2_{U_0}}{\text{Var}(x_i) * (1 - R^2_X) * (T - 1)}}$$

$$\text{WP effect } SE_{\beta_x} = \sqrt{\frac{\sigma^2_e}{\text{Var}(x_i) * (1 - R^2_X) * (T - 1)}} \quad (3.8)$$

in which the square root is used to transform the sampling variance of the fixed effect into a SE metric (i.e., from a variance to a standard deviation metric). The denominator is based on the variance in the predictor x_i that is not accounted for by the other predictors and the product of *T* total observations minus 1; thus, to the

extent that x_i has more variance in general (or less shared variance with other predictors), the SE for its effect will be smaller. Here, though, the numerator includes the remaining outcome variance that is relevant for assessing the strength of the predictor's effect, but which sources of variance are relevant will depend on the source of the effect. The SE for all effects will include at least σ^2_e, but the SE for any BP effect will also include $\tau^2_{U_0}$. This difference is the source of the differing SEs for the WP effects in Table 3.1.

Let us first examine the SE for the simple main effect of group (either at pre-test or post-test) in both models for $T = 100$ total observations (50 persons by two occasions). Because the binary group variable has a mean of 0.50 (i.e., half of the sample is in each group), its variance is 0.5*(1 – 0.5) = 0.25. Given the coding of time (0 = pre-test, 1 = post-test) and group (0 = control, 1 = treatment), the variables for group and the time by group interaction are correlated, such that half of the variance in the group variable is predicted by the interaction ($1 - R^2_X = .50$). Using these values, the denominator of the SE equation will be 12.50 in either model. Critically, because the simple effect of group accounts for BP variance, the numerator is calculated as $\tau^2_{U_0} + \sigma^2_e = 22.78 + 4.45 = 27.22$. The square root of 27.22 / 12.5 = 1.48, which is the SE for the group effect at either occasion. Because the SE for the group effect is calculated using the total remaining outcome variance, the SE is the same in the WP model (in which the remaining outcome variance has been partitioned into two sources of $\tau^2_{U_0}$ and σ^2_e) as in the BP model (in which all remaining variance is contained within σ^2_e). The significance of the group effect can then be determined by a Wald test, in which the slope estimate is divided by its SE to form a *t*-statistic, which is then compared to a *t*-distribution with denominator degrees of freedom equal to *N* persons minus *k* fixed effects. In both models, the Wald test returns a *t*-statistic > 1.96 at the post-test occasion only, and thus the effect of group is significant at post-test but not at pre-test, as found in both models.

Next, let us consider the SE for the simple main effect of time (for either group). The binary time variable also has a variance of 0.25, of which 50% is predicted by the time by group interaction, resulting in a denominator of the SE equation of 12.50 in either model. In contrast, the numerator of the SE equation will differ across models. Because the effect of time operates within persons, its SE is calculated using only σ^2_e, which is defined differently across models, as illustrated in Figure 3.3. In the BP model (top panel), $\sigma^2_e = 27.22$ and contains *all remaining outcome variation:* e_{ti} is the difference between the observed y_{ti} outcome and the $\hat{y}_{ti}$ outcome predicted by time, group, and their interaction (as shown by the dashed line for each group). In contrast, in the WP model (bottom panel), $\sigma^2_e = 4.45$ and contains only the remaining y_{ti} variation relative to the dashed line for each person created from $\hat{y}_{ti} + U_{0i}$ (i.e., after taking into account constant person mean differences). Because of this difference in σ^2_e, the fixed effect of time is tested against less variance in the WP model, reducing its SE from 1.48 in the BP model to 0.60 in the WP model. Its resulting Wald test is significant for both groups in both models, however.

Finally, let us examine the interaction of time by group, whose SE also includes σ^2_e only, and thus differs from 2.09 in the BP model to 0.84 in the WP model. As a

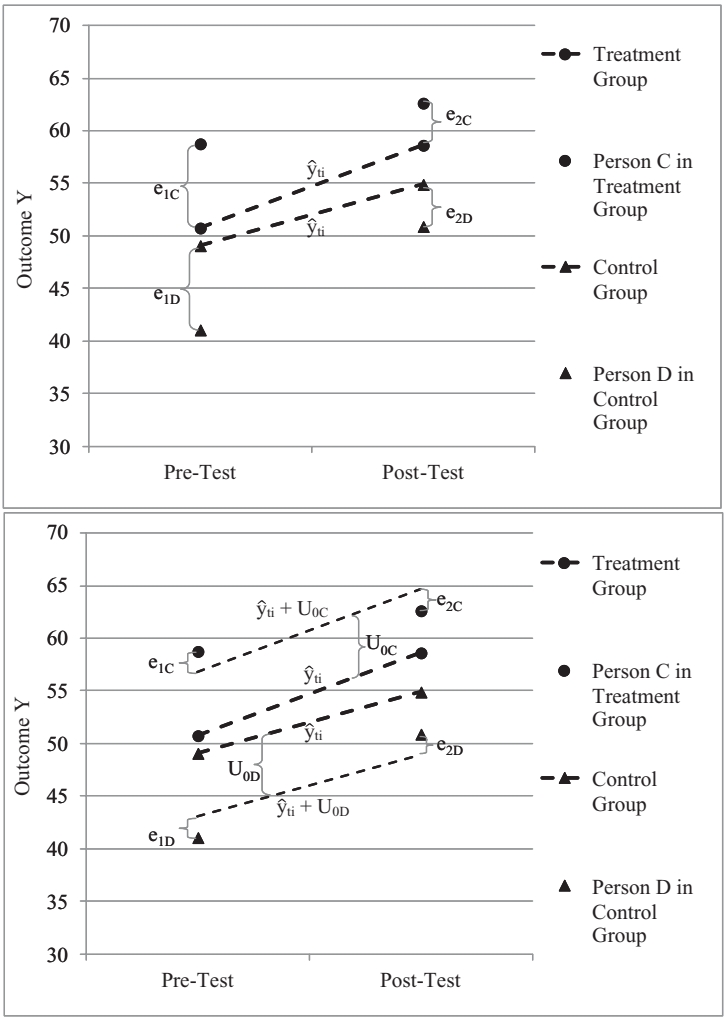

Figure 3.3 Illustration of conditional between-person (top) and within-person (bottom) models for the variance. The data for two hypothetical individuals from different groups are shown.

result, the interaction is not significant in the BP model but is significant in the WP model. So, which set of results should we believe? Although later we will use more formal methods of comparing models, for now we can take a simpler approach. As shown in Equation (3.4), the BP model predicts a correlation of exactly 0 for the model residuals between occasions. We can test this assumption by calculating the e_{ti} model residuals for each observation as: $e_{ti} = y_{ti} - \hat{y}_{ti}$, in which $\hat{y}_{ti} = 49.08 + 5.82(\text{Time}_{ti}) + 1.62(\text{Group}_i) + 2.04(\text{Time}_{ti})(\text{Group}_i)$. The Pearson correlation of the e_{ti} residuals between pre-test and post-test was significantly different from 0 at $r = .84$ ($p < .001$ for $N = 50$ total observations), which strongly suggests that the BP model assumption of 0 residual correlation is incorrect.

In contrast, the WP model predicts a correlation of $r = .84$ for the model residuals between occasions, which can be calculated as the conditional ICC using Equation (3.3) and the estimated $\tau^2_{U_0}$ and σ^2_e variances in the fourth set of columns in Table 3.1 as ICC = 22.78 / (22.78 + 4.45) = 0.84. Because the example data were generated using the WP model in Equation (3.7), in this case the actual residual correlation exactly matches that predicted from the WP model, although this won't always happen. But for now, we can conclude that because the residual correlation is significantly different from 0, the WP model (in which $\tau^2_{U_0}$ predicts a constant correlation of the residuals across time) matches the data better than the BP model (that includes only the σ^2_e residual variance, and thus which predicts 0 correlation of the residuals over time). So, the results from the WP model in which the time by group interaction β_3 was significant are more trustworthy than the results from the BP model—good news for our example researcher!

1.E. Generalizing Results: Fixed and Random Effects

Another way to interpret the results from the WP model is to think of it as a three-way design (with factors of group, time, and person) rather than as a two-way design (with factors of just group and time). In doing so, we will try to identify all possible main effects and interactions and see which our WP model contains, as repeated below from Equation (3.7) in Equation (3.9):

$$\begin{aligned} \text{WP model: } y_{ti} &= \beta_0 + \beta_1(\text{Time}_{ti}) + \beta_2(\text{Group}_i) + \beta_3(\text{Time}_{ti})(\text{Group}_i) \\ &\quad + U_{0i} + e_{ti} \end{aligned} \tag{3.9}$$

in which the main effects of time, group, and person are given by β_1, β_2, and U_{0i}, respectively. Whereas β_1 and β_2 are *fixed* main effects (that belong to the model for the means), U_{0i} is a *random* main effect (that belongs to the model for the variance instead). The difference between fixed and random effects will be elaborated throughout the text, but for now it can be thought of as follows. The **fixed effects** for time and group are used to make inferences about differences in the outcome between *the specific and non-exchangeable variants observed in the study*—here, the pre-test versus post-test occasions and the control versus treatment groups, specifically. In contrast, the **random effect** *is* not *designed to make inferences between the specific variants examined in the study*—we do not care how participant 1 compares with participants 2, 3, or 50, as our sample instead represents a larger population of persons who are seen as exchangeable. We simply want to know if *persons matter*—if there is significant *variance* in the outcome due to systematic mean differences between persons. Because each person provides more than one outcome, we can observe the contribution due to each specific person. Although we could also model differences between persons as fixed effects instead, our current example would require 49 dummy codes to represent all possible differences among the 50 persons. So, rather than model differences between persons with fixed effects, we have modeled differences between persons using the variance of a single random effect U_{0i} that adjusts each

y_{ti} outcome by a constant intercept difference for each person (and we can test if persons matter as whether $\tau^2_{U_0} > 0$). Critically, by treating persons as random we can explore *why* people differ (i.e., explain the $\tau^2_{U_0}$ variance with person characteristics such as group) rather than simply control for those person intercept differences via 49 uninformative fixed effects tied to their ID numbers.

In addition to the three main effects, there are also three possible two-way interactions in our three-way design, as well as one three-way interaction—which of these are included in our model? Although the interaction of time by group is included via the fixed effect β_3, the interaction of person by group cannot be determined, given that each person is in the same group on both occasions. For the same reason, the three-way interaction of time by group by person cannot be determined. That leaves us with the interaction of time by person, reflecting how persons change differently between pre-test and post-test. Although not obvious, the time by person interaction *is* the residual e_{ti}! The only systematic reason left why the observed y_{ti} values do not match those predicted from $\hat{y}_{ti} + U_{0i}$ is because *people change differently over time.* For instance, as shown in Figure 3.3, Person D improved more than was predicted from being in the control group, whereas Person C improved less than was predicted from being in the treatment group. These differences in the time slopes between persons must be considered "error" (as the e_{ti} residuals) given only two occasions, because if we were to give each person his or her own random time slope in addition to a random intercept, the prediction would be perfect! We cannot have a random time slope for each person that is separate from the e_{ti} values without at least three occasions, as we will see starting in chapter 5.

2. Within-Person Models via Repeated Measures Analysis of Variance

The previous example was presented to illustrate the primary difference between between-person and within-person models but was limited to two occasions. This next section addresses traditional models for longitudinal data more generally, in which *within-person models* are more commonly referred to as **repeated measures** (or **within-subjects**) **analysis of variance** (**ANOVA**). In practice, however, there is more than one approach to repeated measures ANOVA resulting from different variants of the basic within-person model, as described next.

To continue, we will use a new example taken from a subset of the *Cognition, Health, and Aging Project* data (as described in chapter 1) in which 101 older adults were measured on six occasions (once per day) over a span of two weeks (with no missing data). The outcome variable was a measure of processing speed—the average response time (RT) across trials (in milliseconds; M = 1,770.70, SD = 494.09, range = 917.67 to 4,159.14) needed to accurately judge whether two series of three numbers were the same or different. The purpose of the study was to assess individual differences in short-term learning; individual trajectories for the

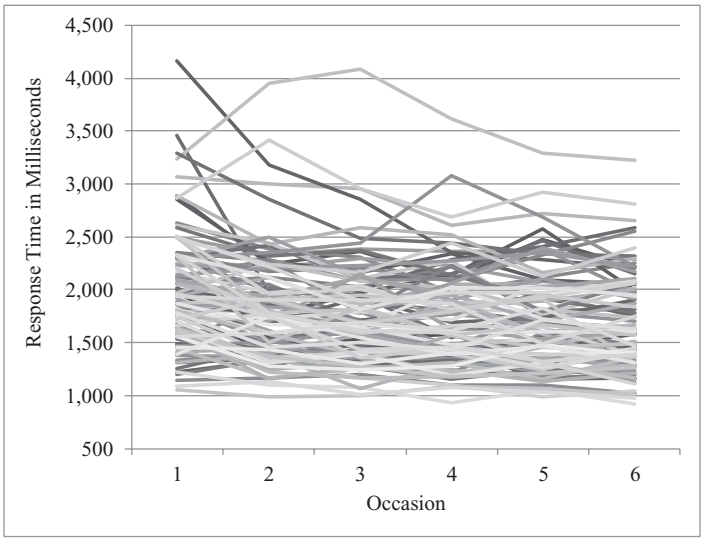

Figure 3.4 Individual trajectories for six-occasion example response time data.

change in RT across the six occasions are shown in Figure 3.4. Descriptive statistics for RT at each occasion are given in the bottom panel of Table 3.2, in which the variance of RT at each occasion is given on the diagonal, the covariance between occasions is given above the diagonal, the correlation between occasions is given below the diagonal, and the means and their SEs at each occasion are given in the bottom rows. Before explaining how three different approaches to repeated measures ANOVA would describe these data (for which the rest of Table 3.2 will be relevant), let us first examine what they have in common—their model for the means over time.

2.A. Saturated Model for the Means in Repeated Measures Analysis of Variance

The repeated measures analysis of variance (ANOVA) model for longitudinal data treats time as a categorical factor. That is, *time* is considered as a set of discrete, fixed conditions in which each person (or individual observational unit: organization, animal, etc.) contributes one outcome variable per time *condition*. The ANOVA model for describing mean change over time uses as many parameters as there are discrete occasions and is thus referred to as a **saturated means model** (i.e., the model for time is *saturated* by using all possible degrees of freedom for differences across conditions of the time variable). A saturated ANOVA model for the means for our six-occasion example response time (RT) data (without yet considering what

Table 3.2 Observed and model-predicted variances (on the diagonal), covariances (above the diagonal), correlations (below the diagonal), means (with SE = standard errors), and fit by model for the six-occasion example response time data.

Between-Person (Independent) ANOVA Model Fit and Predictions Variance Parameters = 1, –2LL = 9155.4, AIC = 9157.4, BIC = 9160.0						
Occasion	1	2	3	4	5	6
1	236,813	0	0	0	0	0
2	0	236,813	0	0	0	0
3	0	0	236,813	0	0	0
4	0	0	0	236,813	0	0
5	0	0	0	0	236,813	0
6	0	0	0	0	0	236,813
Mean	1961.89	1815.17	1750.03	1717.80	1707.18	1672.14
SE	48.42	48.42	48.42	48.42	48.42	48.42
Univariate Within-Person (Repeated Measures) ANOVA Model Fit and Predictions Variance Parameters = 2, –2LL = 8353.4, AIC = 8357.4, BIC = 8362.6						
Occasion	1	2	3	4	5	6
1	236,813	202,677	202,677	202,677	202,677	202,677
2	0.856	236,813	202,677	202,677	202,677	202,677
3	0.856	0.856	236,813	202,677	202,677	202,677
4	0.856	0.856	0.856	236,813	202,677	202,677
5	0.856	0.856	0.856	0.856	236,813	202,677
6	0.856	0.856	0.856	0.856	0.856	236,813
Mean	1961.89	1815.17	1750.03	1717.80	1707.18	1672.14
SE	48.42	48.42	48.42	48.42	48.42	48.42
Multivariate Within-Person (Repeated Measures) ANOVA Model Fit and Predictions (are same as in original data) Variance Parameters = 21, –2LL = 8229.8, AIC = 8271.8, BIC = 8326.7						
Occasion	1	2	3	4	5	6
1	301,985	235,659	217,994	202,607	192,154	195,360
2	0.842	259,150	230,217	213,232	202,092	193,268
3	0.821	0.936	233,368	205,209	196,919	188,604
4	0.791	0.898	0.911	217,544	193,676	185,321
5	0.759	0.862	0.885	0.902	212,098	187,840
6	0.802	0.856	0.880	0.896	0.920	196,733
Mean	1961.89	1815.17	1750.03	1717.80	1707.18	1672.14
SE	54.68	50.65	48.07	46.41	45.83	44.13

kind of model for the variance we will need, thus predicting $\widehat{RT}_{ti}$ rather than RT) is shown in Equation (3.10):

$$\widehat{RT}_{ti} = \beta_0 + \beta_1(T1_{ti}) + \beta_2(T2_{ti}) + \beta_3(T3_{ti}) + \beta_4(T4_{ti}) + \beta_5(T5_{ti}) \quad (3.10)$$

in which five dummy code predictors are used to distinguish the six occasions: T1 = 1 for time 1 and is 0 otherwise, T2 = 1 for time 2 and is 0 otherwise, and so forth. A dummy code for time 6 is not included because time 6 is the reference: β_0 is the mean for time 6, and β_1 through β_5 are the mean differences between time 6 and each other occasion (this coding in which the last category is the reference is the default in SPSS MIXED and SAS MIXED). Because this saturated model for the means contains six fixed effects, it will perfectly reproduce the mean RT at each of the six occasions. Although many other ways of creating contrasts across occasions are also possible, the point is that the default repeated measures ANOVA model for six occasions will include one intercept and five contrasts of some kind. The goal of the ANOVA model is *not* to *predict* or *summarize* the pattern of means via some parsimonious functional form; instead, it simply *reproduces* the observed mean per occasion using fixed effects equal to the number of occasions minus 1 (given that the fixed intercept is already included). This is the case for all repeated measures ANOVA variants that follow—their differences lie not in their (saturated) model for the means, but in their model for the variance.

2.B. Univariate Model for Repeated Measures Analysis of Variance

The univariate model for repeated measures ANOVA is exactly the within-person model for the variance that was described earlier. That is, the univariate model adds two residual terms to the saturated means model in Equation (3.10): the random intercept U_{0i} as the difference between the overall predicted mean (here, $\hat{y}_{ti}$ is just the mean at each occasion) and the mean for person *i* over time (as predicted from $\hat{y}_{ti} + U_{0i}$), and the residual e_{ti} as the time-specific and person-specific deviation of the actual y_{ti} values from those predicted by $\hat{y}_{ti} + U_{0i}$.

Estimating the univariate model for our example data yields a random intercept variance of $\tau^2_{U_0}$ = 202,677 and a residual variance of σ^2_e = 34,316. Through Equation (3.4) we can find the predicted variances and covariances over the six occasions from the univariate model, as shown in the middle panel of Table 3.2. We can evaluate the accuracy of these predictions informally by comparing them with the actual data shown in the bottom panel of Table 3.2. Later in this chapter we will examine a more formal way to compare alternative models for the variance.

First, the univariate model in the middle panel of Table 3.2 predicts the total variance at each occasion to be $\tau^2_{U_0} + \sigma^2_e$ = 202,677 + 34,316 = 236,813, which is exactly the average total variance across the six occasions. As seen in the bottom panel of Table 3.2, however, 236,813 is an underestimate of the total variance for

occasions 1 to 2, but is an overestimate for occasions 3 to 6. Similarly, the univariate model predicts the covariance between occasions to be constant over time at $\tau^2_{U_0}$ = 202,677, which is exactly the average covariance across the 15 possible covariances between occasions 1 to 6. It is an underestimate for 5 of the 15 covariances and is an overestimate for the other 10 covariances. Finally, the **conditional intraclass correlation** (i.e., after including the mean differences between occasions) in the univariate model from Equation (3.3) is predicted to be ICC = $\tau^2_{U_0}$/($\tau^2_{U_0} + \sigma^2_e$) = 202,677 / (202,677 + 34,316) = .86. The correlation between occasions is predicted to result only from the random intercept variance because the model says that the only reason the six outcomes from the same person are related is because of person mean differences (i.e., intercept differences) over time. Because person mean differences are constant over time, the predicted correlation should be constant over time as well. But of the 15 possible correlations between occasions, r = .86 is an underestimate for 9 of the correlations, but is an overestimate for the other 6, indicating that something else is likely causing the correlation over time in addition to constant person mean differences through U_{0i}. But what is the ultimate impact of these discrepancies between the actual data (the variances, covariances, and correlations over time) and the patterns predicted by the univariate repeated measures ANOVA model?

In ANOVA parlance, the pattern of equal variances over time (from $\tau^2_{U_0} + \sigma^2_e$) and equal covariance over time (from $\tau^2_{U_0}$) found in the univariate model is called **compound symmetry**. Considerable research has examined what can happen when compound symmetry does not hold (as in our data). The essential problem is that the SEs for testing mean differences between occasions become biased (and thus so do the *p*-values based on those SEs). Because the model assumes constant variance over time, it uses the same amount of σ^2_e in the numerator of each SE calculation (which may be an overestimate or an underestimate). The result is that for some comparisons the SEs (and thus their accompanying *p*-values) will be too big or too small.

However, as discussed in Maxwell and Delaney (2004, chapter 11), all that is technically needed to obtain unbiased tests of significance of the differences between occasions is for the data to show **sphericity**, otherwise known as the *homogeneity of treatment-difference variances assumption*. Whereas compound symmetry requires equal variances and equal covariances over time, the less restrictive form of sphericity instead requires *equal variance and equal covariance of the pairwise differences between occasions* (e.g., the difference between time 1 and 2, time 2 and 3, and so forth). If the more restrictive assumption of compound symmetry is satisfied, then so is the less restrictive assumption of sphericity. Sphericity is also always satisfied when there are only two occasions, as there is only one possible difference between time 1 and 2 to consider, resulting in a single difference variance and no difference covariances. But for more than two occasions, there are two other variants of repeated measures ANOVA models for addressing violations of sphericity. These approaches will be summarized below; readers who seek greater detail are referred to the excellent treatment by Maxwell and Delaney (2004, chapters 11–14).

2.C. Adjustments to the Univariate Model Tests Based on Degrees of Freedom

Although sphericity tests are available to determine whether the assumption of sphericity has been violated, empirical work has suggested their utility is questionable. Alternative approaches were developed to conduct adjusted significance tests based on a parameter called ε (pronounced "epsilon") that indexes how far off from sphericity the pattern of pairwise variance and covariance over time is determined to be. If sphericity holds perfectly then ε = 1; otherwise ε ranges from 0 to 1, with lower values indicating more deviation from sphericity. Because ε must be estimated from the data, many approaches have been developed to adjust the sample estimate of ε, including the Geisser-Greenhouse lower-bound ε correction, the Huynh-Feldt $\tilde{\varepsilon}$ adjustment, and the Geisser and Greenhouse $\hat{\varepsilon}$ adjustment, the latter of which is recommended by Maxwell and Delaney (2004, p. 545). The goal of each approach is to reduce the degrees of freedom used for the omnibus *F*-test of the overall difference across occasions, creating a more conservative omnibus test. However, the ε adjustment cannot be used to adjust post-hoc comparisons between particular occasions (i.e., at which a single σ_e^2 value may be too big or too small), and an alternative repeated measures ANOVA model is recommended instead, as presented next.

2.D. Multivariate Model for Repeated Measures Analysis of Variance

Although the univariate model (or sphericity-based adjustments thereof) is perhaps the most common repeated measures ANOVA model for longitudinal data, a **multivariate model** could also be used instead. Although both use the same saturated model for the means (i.e., use all possible fixed effects of time to perfectly reproduce the means at each occasion), they differ in what they assume the pattern of variance and covariance over time to be after accounting for all model predictors. The multivariate model cannot be described succinctly by an equation predicting the outcome at any occasion because, unlike the univariate model, it does not include the U_{0i} and e_{ti} terms that imply constant variance over time. Instead, the variance at each occasion and the covariances between occasions are all estimated separately. Another way to say this is that the form of the variance–covariance matrix over time is **unstructured**, meaning that every individual element (variance and covariance) gets to be whatever the data wants it to be. If there are *n* occasions per person, this will result in (*n**[*n* + 1]) / 2 estimated parameters, or for our current 6-occasion example, 6*7 / 2 = 21 estimated parameters (6 variances and 15 covariances).

Because the multivariate model describes the data as they are, the mean differences between occasions are tested more precisely (e.g., the difference between occasions 1 and 2 is tested using their specific variances and covariance, and not the average variance and covariance across all six occasions instead), resulting in SEs and *p*-values for the comparisons that are as accurate as possible. Thus, an

important advantage of the multivariate model is that it can never be wrong—because no assumptions are made about the pattern of variance and covariance across time, concerns about violating sphericity do not apply. A disadvantage of the multivariate model is that the denominator degrees of freedom for these comparisons are based on N persons, and not $T = N * n$ total observations as in the univariate model. Another disadvantage is that the estimation of all possible variances and covariances over time (rather than just $\tau^2_{U_0}$ and σ^2_e) may require large sample sizes, a problem that is compounded as the number of occasions increases.

3. Comparing Alternative Models for Longitudinal Data

So far in this chapter we have seen three model variants for analysis of variance: *between persons, univariate within persons,* and *multivariate within persons.* Although each has the same saturated model for the means over time, they have different models for the variance. So far we have evaluated them informally by observing how well the variances and covariances over time predicted by the model match those observed in the actual data, but there are more formal, empirical ways by which alternative models can be compared. What follows next is intended to introduce the basic concepts and rules for longitudinal model comparison; more technical details will be presented in chapter 5. The indices we will use for model comparison throughout the text are described first, followed by the three primary decision points required to successfully follow the rules of the model comparison process. This section concludes by using these new tools to compare the three ANOVA models and discusses their likely suitability for longitudinal data.

3.A. Introduction to Relative Model Fit Statistics

Although more detail on their origin will be provided in chapter 5, for now we briefly introduce three indices by which **relative model fit** can be judged. The idea of *relative* model fit means that none of these indices can tell us whether a given model fits the data in an absolute sense, but these indices can tell us which of competing models fits relatively better. Indices of absolute fit (i.e., as often provided by structural equation modeling programs by default) are only possible for longitudinal data that have balanced occasions, such that there is a finite set of possible per-occasion means to be predicted by the model for the means, and a finite set of per-occasion variances and covariances to be predicted by the model for the variance.

The three indices of fit we will consider are based on the concept of **model likelihood**, which is an overall summary of the likelihood of observing the y_{ti} outcomes given the estimated model parameters. In likelihood estimation, the program tries out different possible values for the model parameters, and based on the model's assumptions (e.g., multivariate normality of all random effects and residuals), calculates the *likelihood* of observing each y_{ti} value given each set of possible values

for the model parameters. Although similar to the idea of probability, the term *likelihood* is used instead for continuous outcomes. The estimation process terminates (or *converges*) when the next parameter values it tries no longer substantially increase the model likelihood (as indicating by meeting *convergence criteria,* whose values are set by default in the software). The parameter values that result in the highest likelihood are then provided by the program, along with standard errors (SEs) that describe the precision of each parameter estimate. This process of likelihood estimation will be explained in substantially more detail in chapter 5. What is important to remember for now is that through this process we obtain a single value for the likelihood of the model that we will use as the basis of assessing relative model fit.

However, because calculating the likelihood function involves multiplying together the likelihood values for all possible y_{ti} values, the result becomes numerically unstable (i.e., the consequences of rounding error are very large). So, instead of multiplying the likelihood values together to calculate a likelihood function, the natural logarithm of the likelihood function is calculated instead. The **natural log** (abbreviated as just *log* in this text) is the power to which the constant *e* (where $e = 2.718\ldots$) must be raised to return a given number. One of its properties is that by calculating the log of the likelihood function, the likelihood values can then be added together rather than multiplied together, solving the numerical instability problem.

As an end-product of estimation, each model gets a **log-likelihood (LL)** value that reflects how well the estimated model parameters fit the data, which serves as the basis for comparing model fit. The LL value gets multiplied by –2 so that the difference between LL values for two competing models becomes approximately **chi-square (χ^2)** distributed with **degrees of freedom (*df*)** equal to the difference in the number of model parameters. A significance test can then be conducted to see which of two models (if nested; see below) fits better. Accordingly, for many programs (e.g., SAS, SPSS), the overall index of model fit is given not as *likelihood,* but as **–2 log likelihood (–2LL)**. Other programs (M*plus,* STATA) provide the LL values instead of the –2LL values (so that you must multiply those LL values by –2 in order to compare models).

The –2LL value, also referred to as **deviance** in multilevel modeling, will serve as our primary index of relative model fit, in which smaller deviance values (i.e., less positive or more negative) indicate better fit. (Please note that we will refer to the –2LL value *itself* as *deviance* here, even though in other contexts the term *deviance* is sometimes used to describe a *difference* in –2LL values between nested models instead.) However, we will also consider two related indices that take into account *model parsimony,* as shown in Equation (3.11):

$$\begin{aligned} \text{AIC} &= -2\text{LL} + 2 * (\#\,\text{parameters}) \\ \text{BIC} &= -2\text{LL} + \log(\text{N}) * (\#\,\text{parameters}) \end{aligned} \tag{3.11}$$

in which the **AIC** is the **Akaike Information Criterion** and the **BIC** is the **Bayesian Information Criterion** (also known as the **Schwarz Criterion**). Smaller values

indicate a better model for both indices. Often used in this context is the term **parsimony**, which reflects a balance between model complexity and model fit. Although the fit of any model (as indexed by its –2LL value) will be improved by adding more parameters, the improvement should be meaningful in order for the more complex model to be retained. Thus, if two models achieve the same degree of fit, the more parsimonious model with fewer parameters is preferred.

The AIC tries to balance fit (as indexed by the –2LL) with parsimony (via a correction factor of twice the number of model parameters). So, in considering two models with equivalent fit according to the –2LL, the AIC will prefer the more parsimonious model with fewer parameters. The BIC uses a different parsimony correction, the natural log of *N* persons times the number of model parameters. The BIC is more heavily weighted to favor parsimonious models in larger samples, in which the number of parameters will thus count more. Because of these differences in how they correct for parsimony, the AIC and BIC indices may not always agree on which model is relatively best. Nevertheless, these indices of –2LL (deviance), AIC, and BIC will be used throughout the text to compare relative model fit. However, the rules by which they can be used for model comparisons depend on three key decision points, as described next.

3.B. Three Decision Points in Conducting Model Comparisons

First, are the models to be compared **nested** or **non-nested?** That is, can one model be viewed as a subset of the other, such that we only need to add OR remove parameters to get from one model to the other? For instance, if we wished to compare model A whose parameters included main effects of time and group, to model B whose parameters included main effects of time, group, and gender, then model A is *nested* within model B: we can go from model A to model B by just adding gender (or equivalently, go from model B to model A just by removing gender). In contrast, if we wished to compare model A that included effects of time and group to model C that included effects of time and gender instead, then model A and model C would be non-nested: we can only go from model A to model C by adding gender *and* removing group (or equivalently, go from model C to model A by removing gender *and* adding group).

Whether the difference in fit of a nested model is significant can be assessed via the difference in the model –2LL deviance values, or **–2ΔLL** (in which the Greek letter Δ stands for the difference in the –2LL values, pronounced "delta"). This –2ΔLL **deviance difference test** (known more generally as a **likelihood ratio test**, as will be presented in chapter 5) involves three steps: (1) Calculate the deviance difference as: $-2\Delta LL = -2LL_{fewer}$ minus $-2LL_{more}$. The model with *fewer* parameters will have a higher –2LL than the model with *more* parameters, and so the –2ΔLL will be positive. (2) Calculate the difference in the number of model parameters as: $\Delta df = df_{more}$ minus df_{fewer}, so that the Δdf will also be positive. (3) Compare the –2ΔLL deviance difference to a χ^2-distribution with degrees of freedom = Δdf and your chosen alpha level (e.g., $\alpha < .05$). If the –2ΔLL value exceeds the critical χ^2 value for that df, then

the difference in model fit *is* significant: the model with more parameters *fits better* than the model with fewer parameters (or equivalently, the model with fewer parameters *fits worse* than the model with more parameters). If the –2ΔLL value does *not* exceed the critical χ^2 value for that *df,* then the difference in fit is *not* significant: The model with more parameters *does not fit better* than the model with fewer parameters (or equivalently, the model with fewer parameters *does not fit worse* than the model with more parameters). In summary, when relative fit is indicated by the –2ΔLL deviance difference, if adding parameters, model fit can only get *better* or *not better;* if removing parameters, model fit can only get *worse* or *not worse*.

When the models to be compared are *non-nested,* however, no direct significance tests of the difference in their fit are available. Instead, we can compare their AIC and BIC values (given by most programs), such that smaller values indicate better fit. We cannot say the model with the smaller AIC and BIC values fits *significantly better;* instead, we can only say that the model with the smaller AIC and/or BIC *is preferred* given that differences in information criteria do not follow a known distribution and thus no cut-off values are available. Furthermore, unlike the –2LL index of fit, the AIC and BIC indices can indicate worse fit even if model parameters are added, which would mean that the improvement in model fit was not enough to offset the cost from reduced parsimony. AIC and BIC can also be used with nested models as additional evidence that the improvement in fit due to adding parameters is *really* worth it (i.e., ideally, the AIC and BIC will also be smaller if the –2ΔLL test is significant). Thus, the –2LL, AIC, and BIC indices may not necessarily agree on whether a new parameter is helpful to the model.

In addition to whether the models to be compared are nested or non-nested, we must also know exactly *how* they differ from each other. As introduced in chapter 1, all statistical models have two sides: the model for the means and the model for the variance. The model for the means describes how fixed effects of predictors create expected outcome values; the model for the variance describes how model residuals are distributed and related across observations. Thus, a second decision point is, do the models to be compared differ with respect to their model for the means, their model for the variance, or on both sides at once? In the previous comparison example, the models differed only in which fixed effects (e.g., time, group, or gender) were included in the model for the means. As an alternative example, the difference between the univariate and multivariate within-person ANOVA models lies entirely in their model for the variance; both have the same saturated model for the means. As we will see in later chapters, the models to be compared could also differ in both their model for the means and for the variance.

Finally, after determining whether the models to be compared are nested or non-nested and exactly how they differ, we must consider which estimator has been used: **maximum likelihood (ML)** or **restricted (residual) maximum likelihood (REML).** As explained in more detail in chapter 5, ML maximizes the likelihood of the *full data,* treating the fixed effects as *known,* whereas REML maximizes the likelihood of the *residuals* only, treating the fixed effects as *unknown.* These two differences in how the ML and REML likelihoods are computed lead to an important difference in which aspects of model fit will be indexed by their –2LL, AIC, and

BIC values. When using ML estimation, the –2LL, AIC, and BIC indices describe the fit of the *entire model*. As a result, models that differ on either side—in their model for the means and/or in their model for the variance—can be compared using the –2LL, AIC, and BIC indices from ML. In contrast, these indices in REML estimation only describe the fit of the *model for the variance,* and thus can only be used to compare models that differ in their random effects or residual variance parameters. This is because REML maximizes the likelihood of the model residuals specifically—and because the residuals from models with different fixed effects are defined differently, the REML likelihoods based on these different definitions of residuals will not be on the same comparable scale. Thus, the –2LL, AIC, and BIC indices from REML *cannot* be used to indicate if adding new fixed effects in the model for the means has improved model fit. This is not a serious limitation to the use of REML estimation, however, because the significance of each fixed effect can each be assessed more directly via its Wald test *p*-value (i.e., based on its estimate/SE, as we have been doing so far) and multiple fixed effects can be tested simultaneously using multivariate Wald tests (as discussed in chapter 5) within ML or REML.

Given that the REML fit indices can only be used to compare different models for the variance, you might be wondering why anyone would ever use REML estimation in the first place. The reason is that ML has a downside—as stated earlier, ML does not take into account the uncertainty from estimating the model fixed effects, whereas REML does. As a result, REML estimates of variances will be correct, but ML estimates of variances will be downwardly biased (too small) by a factor of $(N - k) / N$, where N is the number of persons and k is the number of fixed effects. This bias in the estimated variances will propagate to create SEs for the fixed effects that are too small under ML as well. However, the difference in the variances estimated in ML or REML will diminish as the number of people increases, and so it will be more important if you are analyzing smaller samples to use REML than if you are analyzing larger samples (in which the bias in ML estimates may be negligible). On that note, be careful to check which estimator is being invoked: REML is the default in some programs (SAS, SPSS, and STATA), but not in others (M*plus*). When using REML, the –2LL value will usually be provided as *–2 res log likelihood* (or something similar), rather than *–2 log likelihood* as provided when using ML.

3.C. Comparisons of Analysis of Variance Models for Longitudinal Data

Now that the rules of model comparisons have been introduced, let us formally compare the three potential ANOVA models (between-persons, univariate within-persons, multivariate within-persons) for our six-occasion example response time data. We will also consider the pattern of variance and covariance they predict over time and their resulting implied patterns of longitudinal change. Although the results reported previously were from REML estimation, because these models all contain the same saturated means model, we can compare the fit of their different variance models using their –2LL, AIC, and BIC values, as given for each model in Table 3.2. Furthermore, recall that the actual data (means and their SEs, variances,

covariances, and correlations across occasions) provided in the bottom of Table 3.2 can be considered the correct answer for what these models are trying to predict.

Our first alternative, the between-persons (BP) ANOVA model (in the top panel of Table 3.2) includes only a single residual e_{ti} (with constant predicted variance across occasions of $\sigma^2_e = 236{,}813$) that captures all the deviations between each actual y_{ti} and the $\hat{y}_{ti}$ values predicted by the model for the means. The BP model predicts zero covariance between occasions—that the residuals from the same person will be no more related than those from different persons. Although this independence assumption is unlikely to be met in longitudinal data, it is actually an empirical question that we'll answer shortly. Furthermore, in terms of the model for the means, while the means at each occasion are perfectly reproduced, their SEs are not—the BP model predicts them to be constant across occasions (because the variance over time is predicted to be constant), whereas in the original data the SEs decline steadily across occasions instead (because the variance in the original data declines over time as well). The omnibus *F*-test for the mean differences across occasions is reported as $F(5, 500) = 4.73$, $p < .001$. But given the misfit of the BP model to the variances and covariances over time, should we believe it?

Our second alternative, the univariate within-persons (WP) ANOVA model (in the middle panel of Table 3.2) includes two residual terms: (1) the random intercept U_{0i} (with a variance of $\tau^2_{U_0}$) to represent deviations from the predicted $\hat{y}_{ti}$ values to each person's mean (then given by $\hat{y}_{ti} + U_{0i}$), and (2) the residual e_{ti} (with a variance still denoted as σ^2_e) that now represents the remaining deviation from $\hat{y}_{ti} + U_{0i}$ to each actual y_{ti}. The total variance is still predicted to be constant across occasions at $\tau^2_{U_0} + \sigma^2_e = 202{,}677 + 34{,}316 = 236{,}813$. Furthermore, just as in the BP model, the means at each occasion are perfectly reproduced by the univariate WP model, with predicted SEs that are again constant across occasions (because the variance is still predicted to be constant). The omnibus *F*-test for the mean differences across occasions from the univariate WP model is $F(5, 500) = 32.85$, $p < .001$. This *F*-value is much larger than in the BP model because 86% of the original remaining variance was moved into $\tau^2_{U_0}$, resulting in much less within-person error (σ^2_e) with which to test the mean differences across occasions.

This partitioning of the outcome variance into $\tau^2_{U_0}$ and σ^2_e also serves an important role in modifying the predicted covariance across occasions. Rather than assuming *no* covariance over time as in the BP model, the univariate (*compound symmetry*) WP model assumes a *constant* covariance over time that is due entirely to the random intercept variance $\tau^2_{U_0} = 202{,}677$ (as seen in the upper off-diagonal in the middle panel of Table 3.2). What this says is that the only reason why the e_{ti} residuals from the same person were correlated originally is because that person was simply higher or lower than the rest of the sample at every occasion. After incorporating each person's mean difference into the model via the U_{0i} random intercept, the e_{ti} residuals are no longer correlated (i.e., independence holds for the e_{ti} residuals only after accounting for U_{0i}). What this implies less directly is that *everyone changes the same*—the only way that people are allowed to differ from each other is in their intercept. As will be discussed in chapter 5, though, if people change differently over time, then the variance over time cannot be constant, and

the covariance or correlation over time cannot be constant, either. Thus, to the extent that individual differences in change are observed, the univariate WP model is not likely to describe the data sufficiently (and creating more conservative omnibus tests by adjusting the degrees of freedom based on deviations from sphericity does not solve this fundamental problem).

So which is the better model for these data: the BP model (with σ^2_e only) or the univariate WP model (with $\tau^2_{U_0}$ and σ^2_e)? Because the BP model is nested within the univariate WP model (i.e., they differ by $\tau^2_{U_0}$), we can use the 3-step likelihood ratio test that was described earlier (i.e., calculate the deviance difference, calculate the df difference, and then compare against the χ^2-distribution). For step 1, the $-2\Delta LL = 9155.4 - 8353.4 = 802.0$. For step 2, the $\Delta df = 2 - 1 = 1$. For step 3, the critical χ^2 value for df = 1 at $\alpha < .05$ is 3.84, which is much smaller than the obtained $-2\Delta LL = 802.0$. We can obtain the exact *p*-value for the obtained $-2\Delta LL = 802.0$ for $df = 1$ by using the χ^2 function in Microsoft Excel (or other programs) as $p = 1.98E^{-176}$ (a number so small it has 176 zeros after the decimal, which we will summarize as $p < .001$). The likelihood ratio test results would be written like this, with the number of parameters added between models as the df in parentheses: $-2\Delta LL(1) = 802.0$, $p < .001$. Equivalently, it could also be written as: $\Delta\chi^2_1 = 802.1$, $p < .001$, explicitly recognizing that $-2\Delta LL$ value is χ^2-distributed with one degree of freedom here.

One complication that you should be aware of when using $-2\Delta LL$ tests (as will be elaborated in chapter 5) is that the $-2\Delta LL$ is only χ^2-distributed with $df = \Delta df$ for the number of added parameters when those added parameters do not have a boundary. This means that the $-2\Delta LL$ is χ^2-distributed with $df = \Delta df$ when adding new fixed effects (i.e., that could be estimated as any positive or negative value), but not when adding a random intercept variance (which can only be 0 or greater and implies a positive average covariance across time). Instead, the $-2\Delta LL$ when adding a random intercept variance is distributed as a mixture of the χ^2-distributions for $df = 0$ (for the missing negative side of the sampling distribution for the random intercept variance) and for df = 1 (for the observed positive side of its sampling distribution).

Several remedies to this problem have been proposed; a simple solution that works when testing a new random intercept variance as we've done so far is to use a one-tailed test (or $\alpha < .10$) instead of a two-tailed test ($\alpha < .05$), which would result in a critical χ^2 value for $df = 1$ of 2.71 instead of 3.84. Unfortunately, this simple solution may not always apply when testing the differences between more complex variance models, as will be discussed in chapter 5. For now, though, we note that the traditional $-2\Delta LL$ test (in which the χ^2 $df = \Delta df$ for the number of added parameters) is overly conservative when used to test new random effects variances. As such, any significant result can be accepted without concern, whereas the influence of using the incorrect χ^2-distribution should be considered in evaluating any nonsignificant results. In our current example, the $-2\Delta LL = 802.0$ would be significant according to either χ^2 critical value (the more conservative but approximately correct critical value of 3.84 from assuming $df = 1$, or the more correct critical value of 2.71 assuming that $df = 1$ only 50% of the time), so the boundary issue is moot. But from this point forward we will explicitly acknowledge the conservative nature of the $-2\Delta LL$ test when relevant (i.e., for the addition of new random effects variances, in which

the Δdf is used as the χ^2 *df* naively without regard to the boundary problem) by including a ~ with the *df*. So, for example, our comparison of the BP model (σ^2_e only) to the univariate WP model ($\tau^2_{U_0}$ and σ^2_e) model would be written as: $-2\Delta LL(\sim 1) = 802.0$, $p < .001$.

The significant $-2\Delta LL$ value indicates that the univariate WP model (with $\tau^2_{U_0}$ and σ^2_e) fits significantly better than the BP model (with σ^2_e only), or more directly, that $\tau^2_{U_0}$ is significantly larger than 0. The smaller AIC and BIC values from the univariate WP model also support it as the better model. The *conditional intraclass correlation* we calculated previously as ICC = $\tau^2_{U_0}$ / ($\tau^2_{U_0}$ + σ^2_e) = 202,677 / (202,677 + 34,316) = .86 (as seen in the lower off-diagonal in the middle of Table 3.2) provides an effect size for this comparison. So, we know that after controlling for occasion mean differences, the outcome contains 86% between-person (random intercept) variance, and that .86 is indeed significantly larger than 0. Thus, the univariate WP model results are more trustworthy than the BP model results. But is the univariate WP model (that predicts constant variance and constant covariance across occasions) good enough *per se*?

As a final point of comparison, we turn to the multivariate WP ANOVA model (reported in the bottom of Table 3.2). Unlike the univariate model that tries to predict the variances and covariances using just $\tau^2_{U_0}$ and σ^2_e, the *unstructured* multivariate model estimates each variance and covariance over time separately, and uses those separate estimates in testing fixed effects. This is why the SEs for the means across occasions are exactly right—because the variances and covariances predicted by the multivariate model are exactly right. The omnibus *F*-test for the mean differences across occasions from the multivariate model is $F(5, 100) = 16.72$, $p < .001$, a more conservative result than the *F*-test returned by the univariate model.

To see if the 21-parameter multivariate model that uses all possible degrees of freedom to reproduce the 21 variances and covariances over time is "better enough" than the two-parameter univariate model that uses just $\tau^2_{U_0}$ and σ^2_e (which is thus nested within the multivariate model), we can compare their fit via a likelihood ratio test. To conduct our test, we calculate $-2\Delta LL = 8353.4 - 8229.8 = 123.6$ and $\Delta df = 21 - 2 = 19$. In this case, though, we don't need to worry about the validity of the $-2\Delta LL$ test because the extra variances and covariances can be thought of as time-specific deviations from the average variance and covariance over time that were given by the $\tau^2_{U_0}$ and σ^2_e model, and those deviations are not bounded. The critical χ^2 value for $df = 19$ at $\alpha < .05$ is 30.14, which is smaller than the obtained $-2\Delta LL = 123.6$ (with an exact *p*-value of $2.35E^{-17}$). So, the multivariate model fits significantly better than the univariate model, $-2\Delta LL(19) = 123.6$, $p < .001$, and the smaller AIC and BIC values for the multivariate model concur. Because it will always fit perfectly, the only real question is whether the improvement in fit of the multivariate model is enough to justify its extra parameters (i.e., all possible variances and covariances over time, rather than just $\tau^2_{U_0}$ and σ^2_e). On this basis you might conclude that the multivariate WP ANOVA model should always be used for longitudinal data. Unfortunately, though, it has two data requirements (which are common to the BP and univariate WP ANOVA models as well) that can severely limit its usefulness for longitudinal data in practice.

First, because ANOVA results are obtained via least squares estimation, complete data are required—persons who miss just one of the time conditions will have *all* their observations dropped from analysis (i.e., listwise deletion). Second, all ANOVA models require that time is balanced across persons—that *time* is composed of a set of common, discrete conditions. As discussed in chapter 1, this requirement will not be satisfied when persons are not measured at exactly the same occasions. For instance, in the current example, *occasion* was our index of time under the assumption that it only mattered how many times during the two weeks participants had practiced the test; accordingly, everyone had the same balanced occasions of 1 to 6. In contrast, if we thought the number of days that had passed between occasions was more relevant, then we could have used *day* as our index of time instead, in which each person would have had a distinct set of values corresponding to his or her time observations during the two-week period. But then the multivariate model with a separate mean and variance at each occasion would not have been estimable because no one would have had observations for each of the 14 days.

Although these are serious practical limitations to the use of WP ANOVA models for longitudinal data, the predictions that they make (e.g., *compound symmetry* or *unstructured* models for the variance) will re-appear as special cases of the models to be presented (but in which listwise deletion will no longer be required). So, ANOVA models are still useful to understand before moving forward to more complex longitudinal models in the next chapters.

4. Chapter Summary

This chapter covered three topics with respect to models for the variance in longitudinal data. First, this chapter introduced the need to distinguish variation in a longitudinal outcome that is *between persons* (BP) from variation that is *within persons* (WP) over time. So far, the only kind of BP variation we have examined is in the mean outcome over time, as represented by a random intercept U_{0i} (and whose variance across persons $\tau^2_{U_0}$ is the estimated model parameter rather than the individual random U_{0i} values). The remaining within-person deviations between the actual y_{ti} outcomes and the person's predicted outcomes (from $\hat{y}_{ti} + U_{0i}$) are then represented by the residual e_{ti}, whose variance across occasions and persons is estimated as σ^2_e. Although no additional outcome variance is explained through this process of partitioning the outcome variation into $\tau^2_{U_0}$ and σ^2_e, it does change the standard errors for the fixed effects in the model for the means, as was demonstrated using an example of treatment and control groups measured at pre-test and post-test. Specifically, we saw that effects targeting BP variation will be tested using both $\tau^2_{U_0}$ and σ^2_e, whereas effects targeting WP variation will be tested using only σ^2_e.

Second, this chapter presented analysis of variance (ANOVA) models for longitudinal data in terms of their implied model for the variance (using a working example of improvement in response time over six occasions). The BP (i.e., between-groups, σ^2_e only) ANOVA model predicts constant variance but no covariance over time whatsoever. In contrast, the WP (repeated measures) ANOVA models do predict

some covariance over time, but the univariate and multivariate versions do so differently. The univariate WP ANOVA model predicts constant variance over time as $\tau^2_{U_0} + \sigma^2_e$ and constant covariance over time as $\tau^2_{U_0}$, a pattern known as *compound symmetry*. Technically, only the less restrictive form of *sphericity* (constant variance and covariance of the pairwise differences between occasions) is required for the univariate WP ANOVA results to be accurate. Although adjustments have been proposed (based on deviations from sphericity) to the degrees of freedom by which the overall mean differences across time are then tested more conservatively, they don't really solve the problem that compound symmetry often doesn't fit longitudinal data. This is because if people change differently over time, then the variances and covariances have to change over time, too. Even in data showing within-person fluctuation, occasions closer together in time may be more related than occasions further apart. In either case, a compound symmetry model may not be adequate for longitudinal data.

As an alternative, the multivariate or *unstructured* WP ANOVA model does not assume compound symmetry, sphericity, or anything else. In fact, it is not really a model at all—it simply estimates all possible variances and covariances over time as is, and so it can never be wrong. However, the multivariate WP model has some practical limitations. It requires a total of (n*[n + 1]) / 2) estimated parameters for n occasions. And like the other ANOVA models, it requires complete and balanced data (i.e., all persons are measured at exactly the same occasions), and models mean differences between occasions using n – 1 contrasts across the discrete time conditions, rather than trying to summarize the overall trajectory over time.

Finally, this chapter presented some new indices of model fit (–2LL, AIC, and BIC, in which smaller is better for each) and practiced the rules by which they can be used to assess relative model fit. These model fit comparisons will recur frequently in the text, and so it is important to understand three key decision points for their use: whether the models to be compared are nested or non-nested, exactly how they differ (in their model for the means, their model for the variance, or on both sides), and which estimator was used to obtain model parameters—either maximum likelihood (ML) or restricted maximum likelihood (REML). With these answers at hand, the rules for model comparisons can thus be summarized as follows.

First, the difference in fit of nested models can be compared via –2LL likelihood ratio tests, in which the –2ΔLL is compared to a χ^2-distribution with degrees of freedom equal to the difference in the number of model parameters. Second, the difference in fit of non-nested models cannot be formally tested, but models with smaller AIC and BIC values are "preferred" (and AIC and BIC can also be used to compare the fit of nested models as well). Third, the fit of *all* nested models can be compared using the –2LL, AIC, and BIC indices from ML estimation, but only different models for the variance can be compared with the –2LL, AIC, and BIC indices from REML estimation. This is because REML maximizes the likelihood of the model residuals rather than the likelihood of the full data as in ML, and so the REML indices from models with different fixed effects in the model for the means are not on the same scale (and will thus not be comparable). But given that Wald tests for fixed effects will provide p-values to assess their significance in both ML and REML, this is not really a problem. Lastly, because it assumes fixed effects are

unknown, REML estimation provides more accurate estimates of variances than does ML estimation by a factor of $(N - k) / N$, where N is the number of persons and k is the number of fixed effects. Thus, it is especially important to use REML rather than ML in smaller samples, which is why we did so in this chapter.

5. Sample Results Sections

The analyses in this chapter could each be summarized into the beginning of a results section as follows (each of which would then need to be expanded to better capture the substantively meaningful story, theoretical framework, or research hypotheses to be tested).

5.A. Two-Occasion Example

The extent to which a new approach to instruction resulted in greater student learning outcomes (M = 53.34, SD = 6.35, range = 37.54 to 68.62) between pre-test and post-test was examined in 50 elementary school children, of which 25 were in a control group and 25 were in the treatment group. A univariate repeated measures analysis of variance model in which persons were treated as a random effect was used to distinguish between-person variation in the mean outcome over time from within-person, time-specific variation. Fixed effects for time (coded such that 0 = pre-test, 1 = post-test), for group (coded such that 0 = control, 1 = treatment), and for a time by group interaction were then examined as shown in the second line of Equation (3.7), and standard errors for the additional model simple effects were obtained via separate syntax statements. Results are shown in the fourth set of columns of Table 3.1; differences in the group means over time are shown in the bottom panel of Figure 3.3 (Note: just the mean trajectories for each group would be shown). The expected outcome for the control group at pre-test was $\beta_0 = 49.08$. The positive effect of time was significant in both the control group ($\beta_1 = 5.82$, SE = 0.60, $p < .001$) and in the treatment group (as given by $\beta_1 + \beta_3 = 5.82 + 2.04 = 7.86$, SE = 0.60, $p < .001$). The higher performance for the treatment group than the control group was not significant at pre-test as expected ($\beta_2 = 1.68$, SE = 1.48, $p = .26$), but was significant at post-test (as given by $\beta_2 + \beta_3 = 1.68 + 2.04 = 3.72$, SE = 1.48, $p = .02$). Finally, the larger improvement over time of the treatment group relative to the control group (the time by group interaction) was significant as expected ($\beta_3 = 2.04$, SE = 0.84, $p = .02$).

5.B. Six-Occasion Example

The extent to which response time (RT in milliseconds) to a measure of processing speed (M = 1,770.70, SD = 494.09, range = 917.67 to 4,159.14) improved over six occasions was examined in a sample of 101 older adults via a saturated means

model (i.e., including five contrasts for mean differences in RT among the six occasions). The extent to which the residual variances and covariances of RT across occasions would be adequately described by three variants of ANOVA models was then examined. As expected, substantial covariance among the residuals from the same person was observed (conditional ICC = .86), as indicated by the significantly better fit of a compound symmetry model (with equal variance and equal covariance over time) than a model with no predicted covariance over time, $-2\Delta LL(\sim 1) = 802.0$, $p < .001$. However, the prediction of equal variance and equal covariance did not adequately describe the actual RT data, as indicated by the significantly better fit of a multivariate (unstructured) model in which each variance and covariance was estimated separately, $-2\Delta LL(19) = 123.6$, $p < .001$. Significant mean differences in RT were observed across occasions within the multivariate model, $F(5, 100) = 16.72$, $p < .001$.

Review Questions

1. Why is it necessary to consider what to include in the model for the variance in longitudinal data? Refer in your answer to the statistical and substantive reasons for doing so.
2. What predictions do the three types of analysis of variance (ANOVA) models (between-persons, univariate within-persons, and multivariate within-persons) make about the pattern of means, variances, and covariances over time for a longitudinal outcome? What are the limitations of these ANOVA models for longitudinal data?
3. How are the –2LL, AIC, and BIC indices used to compare relative fit? What kinds of model comparisons can be made using these indices when using ML versus REML?

References

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data*. Mahwah, NJ: Erlbaum.

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

Snijders, T.A.B., & Bosker, R. (2012). *Multilevel analysis* (2nd ed). Thousand Oaks, CA: Sage.