22 Dealing with Repeated Measures

Design Decisions and Analytic Strategies for Over-Time Data*

Amie M. Gordon and Katherine R. Thorson

22.1 Introduction

Over-time, longitudinal, and repeated measures are all terms referring to data with repeated measurements of the *same variables within the same unit* (e.g., person, family, team, company). At a minimum, data may have only two timepoints (e.g., measuring pre- and post-intervention), but they can also be much more intensive (e.g., second-by-second measurements within a social interaction). When people think about longitudinal data, self-report data often come to mind, but longitudinal data can be observational, behavioral, or physiological as well. Smartphone use, health-based data, physiological responses, or academic outcomes can all be longitudinal if you collect the same data repeatedly from the same unit.

22.1.1 Why Do Researchers Use Longitudinal Methods?

Researchers use longitudinal methods for different reasons. One prominent reason is because they are interested in modeling change over time (e.g., how do feelings of belonging change across the four years of college?). Another reason is because they want to know whether the association between two variables exists within a person (a within-person effect) or across people (a between-person effect; Gable & Reis, 1999). For example, if examining stress and well-being, a within-person effect would focus on whether people experience lower well-being on days when they are more stressed compared to days when they are less stressed, whereas a between-person effect would focus on whether people who are more stressed tend to experience lower well-being compared to people who are less stressed. A third common reason for collecting repeated-measures data is to increase measurement reliability. For instance, instead of asking people to report about their personality at one timepoint, researchers may ask participants to report about their personality on many different days to try to obtain more stable, reliable estimates.

In writing this chapter, we had three main goals:

1 help researchers consider design decisions when developing a longitudinal study,
2 describe the different decisions researchers have to make when analyzing longitudinal data, and
3 consider the unique properties of longitudinal designs that researchers should be aware of when designing and analyzing longitudinal studies.

This chapter is not meant to answer every question about longitudinal data. Instead, we aim to provide a comprehensive overview of the major issues that researchers should consider, and we also point to more extensive resources.

* We wish to thank Niall Bolger, David Kenny, Harry Reis, Tessa West, C. J. Concepcion, Emily Diamond, Annika From, and Micaela Rodriguez for their helpful feedback on a prior version of this chapter.

22.2 Design Decisions

There are several common types of repeated-measures design (see Table 22.1) and many aspects to consider when designing a study with repeated measures. For example, one critical design aspect is ensuring that the timescale of your repeated measures matches the timescale of the phenomena you are studying (see Figure 22.1). Ideally, you will choose measurement timepoints that accurately reflect the underlying pattern of temporal change.

Table 22.1 Common types of repeated-measures design

Name	Definition	Examples
Event-contingent	Participants provide data (self-report, behavioral, physiological) in response to a particular event. With advent of technology, also options to make this location-contingent based on location data (e.g, GPS).	Participants are given a link to a survey and are instructed to complete the survey every time they experience a negative interpersonal event. Participants wear an ambulatory heart rate (HR) monitor and are instructed to assess their HR and complete a survey every day when they enter and leave the workplace
Daily diary	Typically a single daily assessment	Participants are sent a link to a survey each night for two weeks and told to complete the survey right before bed. Participants are sent a link to a survey each morning for a week and asked to complete the survey right after they wake up.
Experience sampling method (ESM) Ecological momentary assessment (EMA)	Typically multiple times a day with the goal of capturing a snapshot of natural life. Typically multiple times a day with the goal of capturing momentary experiences. Practically EMA and ESM are used interchangeably for methods that use multiple assessments per day for a number of days	For one week, participants are sent text messages at random times throughout the day with a link to a survey asking them to report on their social interactions in the past 30 minutes. Participants download a research app that prompts them to complete surveys and physiological measures, such as HR and blood pressure (BP), every morning, afternoon, and evening for three weeks.
Pre–post	Gathering data before and after an event (can be a naturally occurring event or some type of manipulation or intervention)	Participants complete surveys and behavioral and physiological tasks in the lab before and after a one-month online intervention. Participants are sent links to surveys before and one week after an election.

Table 22.1 (Cont.)

Name	Definition	Examples
Longitudinal	Although all of these designs are longitudinal, in practice this term is often used to refer to designs with fewer repeated measures that span longer periods of time and are collected with longer time intervals between them, often across months or years (as opposed to "intensive longitudinal designs," like daily diary or ESM studies that typically include frequent repeated measures at close intervals)	Participants who just got married are sent a link to a survey every six months for three years. Participants who started college are brought into the lab to complete surveys and behavioral and physiological tasks at the beginning and end of each academic year.

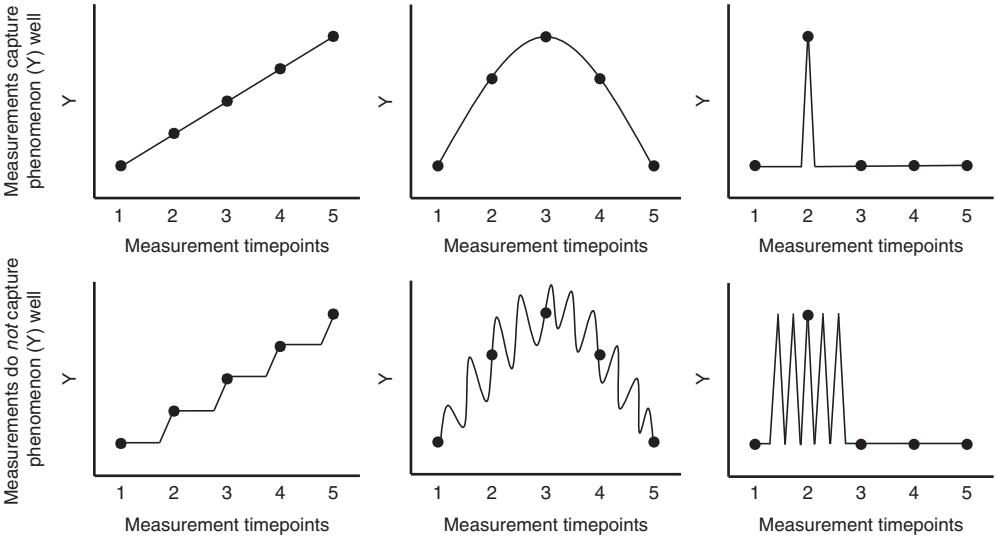

Figure 22.1 Figures display patterns of change in a phenomenon (*Y*) over time. Dots represent values of *Y* obtained by measurements at five timepoints. Within the same column, the data obtained in rows 1 and 2 are the same. However, in the first row, the measurement timepoints capture the underlying pattern of change; in the second row, the measurement timepoints miss meaningful information about the underlying pattern of change

As you plan your study, we encourage you to think through the questions you want to answer and the analyses you will eventually run as much as possible. As we both learned firsthand, sometimes a few tweaks in a design can save you a headache at the analytic stage.

22.2.1 Frequency and Timing of Repeated Measures

One decision you have to make when designing a longitudinal study is how frequently the data will be collected, as well as the distance from initial to final data collection. At the forefront of your decision making should be the purpose of collecting data over time: what questions are you trying to answer?

22.2.1.1 Equal versus Unequal Spacing

If you are examining change across timepoints, equal spacing between repeated measures (e.g., collecting data every four months – four, eight, and twelve months after an event of interest) typically allows for a more parsimonious analytic model than unequal spacing. However, you may have theoretical or practical reasons for unequal spacing. For instance, you may not be interested in predicting change across many timepoints or from one timepoint to the next, but rather in looking at the effects of baseline processes (i.e., variables collected at an initial measurement) on short- and long-term changes. One of us conducted a study tracking couples one month, six months, and twelve months after a baseline lab session so that we could see how baseline processes predicted change at shorter and longer follow-ups. Researchers can also face fiscal constraints that make equally spaced timepoints difficult. If a researcher wants to study reactions to an election and does not have funds to collect twelve weeks of data, they might collect data one week before the election, right after it, and three months later. With projects like these, the question is less about tracking consistent change over time and more about reactions before, during, or after a particular event of interest. Researchers may also end up with unequal spacing due to the nature of their variables. A researcher who is interested in work experiences might conduct a daily diary study for two weeks on Monday through Friday, creating unequal spacing due to weekends. Weekends can be marked as missing data points if the researchers want to model time continuously, but they are not missing at random and could be meaningfully different from the rest of the data. Lastly, researchers might also use an event-contingent design in which participants are instructed to complete a participant-initiated survey each time an event occurs (e.g., the Rochester interaction record; Reis & Wheeler, 1991; see Reis, Sels, & Gable, Chapter 12 in this volume). This type of design typically creates unequal spacing because the event of interest does not occur in an equally spaced manner (e.g., if participants report every time an interpersonal conflict occurs, this might be three times a day for some people and once a week for others).

22.2.1.2 Frequency of Variables of Interest

As noted above, an important aspect to consider is how often the processes or behaviors you are interested in occur. Some processes, like relationship conflict or discrimination, tend not to occur on a daily basis (e.g., Gordon & Chen, 2014; Harris et al., 2022), making it difficult to capture them within a shorter time frame. Collecting data daily for a week in this case may yield less useful information than collecting data once every few days or once a week for a longer period of time.

22.2.1.3 Stability of Variables of Interest

How frequently do the processes you are studying change? If measurements only occur during a time window in which little change occurs (e.g., relationship satisfaction during the honeymoon phase), you may end up collecting a lot of data with little variability (this stability may be useful information, but researchers collecting repeated-measures data are often interested in modeling change). If your variable of interest is slow-moving, you may want to spread your measurements out over a longer period of time. For example,

heart rate changes rapidly and you may assess change every few seconds, whereas salivary cortisol assessments are typically spaced further apart, given slower changes (Thorson et al., 2018). This is true with psychological and behavioral processes as well. For example, we have found that people show more day-to-day variation in relationship conflict than they do in relationship satisfaction (Gordon, 2023).

22.2.1.4 Proximity of Data Collection to the Event of Interest

If your study is capturing information about specific events or behaviors, consider how close data collection needs to be to the event or behavior. If you want to capture data about the event quickly (e.g., emotion regulation right after an exam), you might want to use an event-contingent design. The less you care about capturing the event as it occurs, the further apart you can space your timepoints. Sometimes a researcher is interested in information about an event or experience that is easy to recall (e.g., rare and memorable events such as whether a couple broke up but not details about the psychological processes that occurred during those events). In these situations, conducting a study with longer spacing between follow-up surveys may be preferable.

22.2.1.5 Timeframe of Variables

For designs that are not event-based, consider the timeframe in which you want to capture an experience. Do you want to capture people's emotions at a given moment or their mood across an entire day? Thinking through the operationalization of your variables (e.g., what are the exact self-report questions that you will ask?) can help you decide, for example, whether you need to capture data multiple times a day or whether only once a day will suffice (Chun, 2016). Here, again, it is useful to think about the timescale of the phenomena you are studying (see Figure 22.1). If a process fluctuates throughout the day – emotional experiences, for example – and you want to capture those fluctuations, you might ask about it several times a day and in reference to the past thirty minutes (e.g., "How excited have you felt over the past thirty minutes?"). If a process is stable throughout the day (but fluctuates from day to day) – certain habits and behaviors, for example – you might ask about it once a day and in reference to the whole day (e.g., "How many minutes did you spend vigorously exercising today?").

22.2.1.6 Minimum Number of Data Points

Often people ask about the minimum number of data points for a repeated-measures study. This will depend on many factors, not least of which is the specific question researchers are asking. For instance, you may be interested in modeling trajectories of variables over time. Pre–post designs – in which measurements are collected before and after some event or process – only require two timepoints. Most designs and questions about change, including questions about within-person variability and the nature of change over time (e.g., whether linear or nonlinear) require many more timepoints, though. For analyses of these types, the more data points the better (see also the subsection 22.2.2 below). In other situations, you may be collecting repeated measures to yield a more reliable estimate of a particular variable or process. In these cases, the number of data points you need will depend, in large part, on the variability of your variables of interest. If the variable is highly stable, one or a few data points may be enough – for example, asking students their GPA weekly will yield little variability. On the other hand, if the variable varies from day to day (e.g., social activities can vary a lot across days), more data points will be necessary to capture a reliable estimate of the average effect.

22.2.1.7 Multiple Frequencies

Think creatively! You can have different kinds of data collected at different frequencies. For example, researchers could track a group of college

students across four years of college and conduct an event-contingent study during exam week at the end of each academic year to examine emotion regulation processes as they relate to stressful academic situations. They could also bring students into the lab at the end of every year to examine changes in emotion regulation in response to an acute stressor. As another example, researchers could collect social interaction data from work teams during a monthly team meeting. At the same time, the researchers could track the teams' quarterly performance and collect quarterly self-reports about team cohesion.

22.2.1.8 Timing of Assessments

You also have to decide exactly *when* your assessments will occur. If you are collecting data once a day, do you want to survey people in the morning, in the afternoon, or at night? Or can people complete the survey whenever they want? This decision should be driven by the questions you are asking. For example, timing would differ if you were interested in capturing participants' expectations for the day rather than having them reflect on the day's events.

If you are collecting data multiple times a day, think about whether you want your reports to be at specific times or randomly distributed. If at specific times, do you want them tied to certain events, like waking, the end of the workday, or bedtime? If randomly distributed, you can consider having multiple assessments that appear randomly within designated time intervals (e.g., sometime between 8 am and 10 am, between 11 am and 1 pm, and so on), to ensure a somewhat even distribution across the day. Also think about the starting timepoint. If you have different questions at different times, starting with the timepoint that holds more of your predictors than outcomes will help maximize the data you can use to answer your primary questions. If you primarily want to use night reports to predict reports the next morning, it makes sense to have participants start at night and end with a morning report, for example.

22.2.1.9 Other Concerns

One drawback associated with frequent check-ins is that reporting on one's experiences can affect people's reports of those experiences (Torre & Lieberman, 2018; see Reis, Sels, & Gable, Chapter 12 in this volume). For example, asking someone five times a day whether they have called their mother may prompt them to call their mother. Or, more seriously, frequently asking someone if they are feeling depressed could cause them to introspect more and change their mood. Reports can also be biased at first such that people initially have stronger responses, and researchers may want to account for this initial elevation bias (Shrout et al., 2018). One way to do this is by including extra "practice" timepoints at the beginning of the study that are collected prior to any particular timepoint of interest and can be excluded from analyses.

22.2.2 Power and Sample Size

In designing your study, you will need to plan sample size. Statistical power calculations are not as straightforward with longitudinal data as with cross-sectional data because you have to decide on sample size for each level of data. If running a daily diary study, you need to decide how many people and how many days (two levels). If you have an ESM study, you have to decide how many times a day as well (three levels). Keep in mind that you will generally increase power more by increasing the number of people rather than the number of repeated measures (Bolger & Laurenceau, 2013; Snijders & Bosker, 2012). Below, we highlight several questions that are important to think through when considering statistical power and sample size.

22.2.2.1 How Similar to Each Other Are Measures from Different Timepoints?

One factor that affects power is the similarity between repeated measures within a person. The more similar repeated measures are, the less unique information you get with each additional

measure. One way to index this similarity is with the intraclass correlation coefficient (ICC; e.g., Schrader et al., 1988; Uhlig et al., 2020), which, in this context, represents the proportion of the total variance in an outcome that is explained by between-person variability in mean levels (for details on calculation, see Chapter 3 of Garson, 2019). To give an extreme example: if you asked adults to report their height each day for a week, you would have all the information you need about each participant after the first day because adult height is unlikely to change from day to day (ICC = 1, or 100 percent of variance is between-person). On the other hand, if you asked adults to use a random number generator and report the number that was generated each day for a week, then there would likely be no correlation between repeated measures within a person, and each new day of data from the same person would be as beneficial (from a statistical-power perspective) as a day of data from another participant (ICC = 0, or 0 percent of variance is between-person). Of course, these are extreme examples, but it's worth knowing that there are meaningful differences in how highly correlated repeated measures can be that depend on the phenomena being measured. Ultimately, the more correlated the repeated measures are within a person, then the less statistical power you have to detect a relationship with another variable that is also collected repeatedly.

22.2.2.2 Are Participants Independent of Each Other?

In addition to thinking about the nonindependence between repeated measures within a person, you must also think about whether there is nonindependence between the people in your sample. If you are studying dyads or groups you may have to account for the correlation between people in the dyads or groups in your statistical models. As with repeated measures, the more similar people from one dyad or group are to each other, the less information you get from each additional person. Returning to our extreme examples, if we had 100 clones who were identical, we would get no additional information from each extra clone. On the other hand, if we had 100 naturally made humans with differing personalities and life experiences, each extra human is likely to add more novel information. More realistically, two siblings are likely to share more personality traits and life experiences than two strangers. In other words, there is less statistical power to find a significant relationship between two variables that are measured in a sample of 200 sibling pairs (400 individuals) than there is to find a significant relationship between two variables measured in a sample of 400 individuals who are not related to each other in any way (see chapter by Kenny, Ackerman, & Kashy, Chapter 23 in this volume, for more information).

22.2.2.3 Do You Care about Between-Person Variability?

Often researchers collecting repeated-measures data are interested in looking at the extent to which people vary from each other (i.e., between-person variability). This might be obvious if your variables of interest are measured only between persons. For example, if you are interested in how motivation at work changes over time and whether this differs for managers versus subordinates, the comparison between managers and subordinates is between persons. However, you might also be collecting variables that include both within- and between-person variability, and it is important to consider which effects you are interested in. For example, you might be interested in associations between sleep and forgiveness. You hypothesize that, on average, people are less forgiving after experiencing conflict on days when they slept worse than they usually do (a within-person effect). However, you might also be interested in whether this is true for everyone or whether people differ from each other in the extent to which their forgiveness is related to how well they slept (between-person differences in the within-person effect).

This between-person question can be tested with longitudinal data, but modeling the between-person heterogeneity requires larger samples than if you only modeled the average within-person effect because the sample size is the number of people in your sample, not the number of repeated measures. Modeling these between-person differences in within-person effects also requires larger samples than would be necessary to model between-person differences in average levels (in our example, that would be whether people who tend to sleep worse than others also tend to be less forgiving after conflict; Bolger & Laurenceau, 2013).

22.2.2.4 How Do I Calculate Power and Effect Sizes to Determine My Sample Size?

Exact equations for calculating power are beyond the scope of this chapter, but we do want to note a few unique considerations that arise when calculating power and effect sizes with longitudinal data.

It is not particularly useful to think of power at the study level. Instead, you have to think about power for each effect that you are interested in, and power for each of these effects might vary depending on the type of effect (e.g., within- or between-person, including random effects or not). If you collect repeated-measures data but will ultimately have only one single data point per participant (e.g., in cases in which you plan to aggregate repeated measures to get a single estimate for each participant or if you are predicting an outcome variable that was measured only once, like GPA at the end of college), then you can use more traditional methods for calculating power (see the subsection 22.3.1 below for more on this). In cases where you are modeling repeated measures within each person and thus have nonindependence and multiple levels of data, proper power calculations are an area of ongoing development. Some resources have been developed for straightforward two-level models in which a single level of repeated measures is nested within individuals (with no additional nesting due to measuring dyads or groups; for example, Monte Carlo simulation methods: Arend & Schäfer, 2019; Bolger et al., 2012; Lane & Hennes, 2018, and the design-effect equation: Hox et al., 2018).

If you have a two-level model (see section 22.3 for more on determining levels in your model), the design effect equation can be useful (Hox et al., 2018; Snijders & Bosker, 2012). This equation uses your total sample size, the average number of repeated measures per person, and the ICC to calculate what is known as the effective sample size (Killip et al., 2004; Snijders & Bosker, 2012). The effective sample size tells you what the sample size would be for a simple random sample with independent observations that has the same precision of estimates (and therefore statistical power) as your two-level sample with nested data. It can help you figure out how large your sample with nested data needs to be to have the same statistical power as a particular sample size with independent datapoints. So you may have 1,000 datapoints from 100 participants providing ten days of data, but if you have a very high ICC, you can think of that as having a sample size that is effectively much smaller than 1,000 (and the design effect equation quantifies what exactly "much smaller" means). To help visualize how varying levels of nonindependence shape effective sample size, Figure 22.2 displays the effective sample size for a study of 100 people with ten repeated measures at varying ICCs.

We refrain from providing links to specific online calculators here to prevent suggesting approaches that may become obsolete or may keep you from looking for more updated programs. However, we encourage you to search for "power calculations for multilevel models" (or try replacing "multilevel models" with "mixed-effects models," "random-effects models," "nested models," or "longitudinal models") to see what is currently available; new R packages are frequently being developed. The better

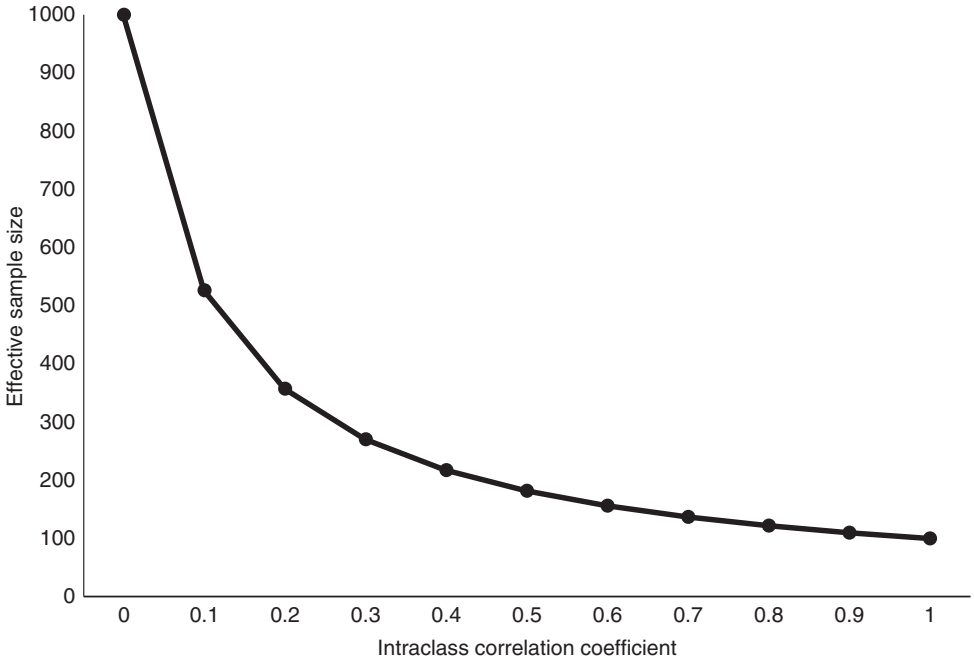

Figure 22.2 Effective sample size at varying levels of within-person similarity in repeated measurements. The figure displays the effective sample size for 100 participants with ten repeated measurements at varying levels of within-person similarity in the measurements (as indexed by the ICC). Higher ICCs indicate more within-person similarity in the measurements

packages and online calculators often accompany a published peer-reviewed article (e.g., see the approach by Lafit et al., 2021). Be aware that a program developed to calculate power in a simple two-level model will not be accurate if your model is more complicated (e.g., more levels of data, such as in an ESM design). If you are using structural equation modeling (SEM), also be aware that power is often calculated for the entire model rather than for each individual parameter estimate.

Effect sizes can also be more complicated to calculate with longitudinal data because you must decide what goes into the denominator, and variance exists at multiple levels. For analytic models that nest repeated measures within people (as opposed to SEM, which typically treats each repeated measure as a separate variable), there are a few papers that provide equations for calculating r and R^2 (Brock & Lawrence, 2008; Edwards et al., 2008; Kashdan & Steger, 2006; Rights & Sterba, 2019). It is important to know that these equations rely on degrees of freedom (dfs), which can also be complicated with repeated-measures data analyzed using multilevel modeling approaches. Unlike most statistical analyses, there are multiple ways of calculating dfs for multilevel models, some of which are fractional and can differ widely between predictors within the same model (i.e., Kenward-Roger dfs or Satterthwaite dfs; McNeish, 2017). These dfs are based on how much of the variance in the outcome is between- versus within-person, and, in a model with repeated measures nested within participants, should range between the total number of repeated measurements and the number of participants, depending on the ICC. These approaches to dfs provide more accurate

estimates than the residual method (also called the within–between method) that calculates dfs as *either* the total number of repeated measurements (for time-varying variables without random effects) *or* the number of participants (for time-invariant variables and time-varying variables with random effects). Different software programs have different default methods for calculating dfs; many programs offer the option to select one of the three methods described above.

22.2.2.5 How Much Missing Data Are You Likely to Have?

You should also think through potential missingness in your data, as this will affect statistical power for a given sample size. You might search for similar past studies to find out whether people tend to drop out at random, or, if not, which factors drive dropouts. For example, are there certain demographic groups that are more likely to have missing data? If so, you may want to oversample from those groups. You can also try to find out, on average, what percentage of missing data you are likely to have based on similar studies.

Note that sometimes missingness occurs because participants drop out of the study completely. To minimize this attrition, you could offer a bonus (e.g., additional pay, a lottery prize) to people who complete at least 80 or 90 percent of the timepoints (at 100 percent people may become disincentivized if they miss a single timepoint; e.g., Foster & Beltz, 2022). Checking in with participants regularly can help maintain engagement (Teague et al., 2018). We have also provided feedback to participants at the end of the study from the data we gathered (i.e., summaries of sleep and stress over time compared to average levels). For long-term studies, sharing results from early waves of data collection might help keep participants motivated (Gordon et al., 2022). You can also plan for a second round of data collection if attrition is high in your first round.

22.2.3 Additional Procedures before Conducting Your Study

As you are planning your study, we also recommend you engage in two processes: (1) conduct a pilot study and (2) think through your study from start to finish.

22.2.3.1 Conducting a Pilot Study

If at all possible, we urge you to run a small pilot study. This can be useful for many reasons, including planning your sample size. By gaining some information about the similarity of repeated measurements to each other (i.e., the amount of nonindependence), as well as effect sizes, you will be able to make more informed estimates of statistical power. Pilot studies can also be a great way to assess issues of timing and spacing, such as the frequency of events of interest (how often do participants report experiencing the event you care about?) and the stability of your variables. If a pilot study is not feasible, you can try reaching out to an expert who studies the processes you are interested in for advice. They might have data you can use to test some of these basic descriptive questions.

22.2.3.2 Thinking through Your Study from Start to Finish

Think through the full lifespan of your study, from start to finish, before launching it to make sure there are no unexpected roadblocks in the way. If you are studying organizational teams, will people participate in a weekly survey over the summer? Are you studying a specific event that means everyone should start the study on the exact same day (e.g., reactions to an election)? If you are conducting a diary or ESM/EMA study, do you care about differences in the day of the week such that everyone should start on the same day of the week? And how many days does the study need to run for? Do you need a full week that includes weekdays and weekends? For example, if you are interested in leisure activities, stress, or socializing with coworkers, these

experiences likely differ on weekdays versus the weekend. You can also think about how long data collection is going to last. What is the expected timing between collecting your first and last participants? Might your variables of interest vary meaningfully during that time? For example, are there expected seasonal patterns (e.g., stress in winter versus summer)? Is there some other change that might affect your results if you are collecting data during that time (e.g., the start or end of school for students and families with children)? If you cannot plan around these events, at least think through collecting the relevant data that will allow you to adjust for them in analyses.

22.3 Statistical-Model Decisions

Once you have collected your longitudinal data, you need to figure out how to analyze them appropriately. The first issue to address is the potential nonindependence between repeated measures within each person.

22.3.1 Nonindependence: What Is It Exactly and Do You Have It?

Nonindependence in repeated-measures data exists when you have potentially correlated errors in your *outcome variable.* That is, you need to account for nonindependence in your data if the variable you are *predicting* has been measured more than once within the same person. This can also be true if it is measured more than once in any unit of analysis such as a dyad, group, or organization, but here we focus on repeated measures within individuals. For example, if you want to predict whether feelings of belonging are associated with mood in daily life and you measure mood daily for a week, then it is likely that one person's mood on Monday is more similar to their own mood on Tuesday than to someone else's mood on Tuesday. In traditional statistical approaches, like a typical OLS regression framework, all predicted data points are assumed to be independent of each other. Correlated repeated measures within the same person violate this assumption, which leads to biased standard errors and degrees of freedom (Fox, 2015).

22.3.1.1 Do You Always Have Nonindependence with Longitudinal Data?

Just because you collected repeated-measures data does not necessarily mean that you are violating assumptions of independence. Because the concern is the outcome variable, it may be that you actually only measured the outcome of interest once per person and thus your data points are independent of each other. One specific situation in which repeated-measures data do not violate assumptions of independence is when you are predicting change in a variable of interest across two timepoints: the outcome variable is the change score or the score at the second timepoint, yielding only one outcome for every person. You may also only have one outcome per person if you are interested in a person-level variable that was only measured once (e.g., predicting end-of-semester GPA from average daily belonging) or you are interested in predicting an aggregated score from repeated-measures data (e.g., number of conflicts in a week; Hox et al., 2018). In these situations, although you have collected repeated measures within a person, you only measure your outcome variable once per person and thus you can use more traditional statistical approaches to analyze your data. You just have to make sure that your *predictors* also reflect the between-person level of analysis. In traditional statistical approaches, you will have biased tests if you predict a score measured once per person from a score measured repeatedly (i.e., have multiple rows of data for the same person, but the outcome variable has the same score on every row). This is known as disaggregation and can lead to tests that are too liberal or too conservative, depending on whether you are testing for between- or

within-person differences (Hox et al., 2018; Snijders & Bosker, 2012).

Often, people are interested in knowing the level of nonindependence that violates assumptions of independence. Does any amount of nonindependence violate the assumption? Or is there a certain amount that means you have too much nonindependence to safely ignore? It's worth knowing that even a small amount of nonindependence can change standard errors in ways that dramatically affect statistical significance (e.g., see Kenny et al., 1998), shifting the conclusions that people make. And, in the case of repeated-measures data, there is almost always some nonindependence over time. Therefore, to avoid making inaccurate conclusions about your data, the safest bet is simply to use statistical models that allow one to adjust for nonindependence, regardless of how much exists.

22.3.2 Dealing with Nonindependence: Identifying the Structure of Your Data

If your outcome variable is measured multiple times within the same person, and there is *any* nonindependence between the repeated measures (which simulation studies suggest is true even for ICCs as small as 0.10; Vajargah & Masoomehnikbakht, 2015) then you must utilize statistical approaches that account for nonindependence. For the remainder of this chapter, we walk you through the main issues to address and decisions to make when conducting analyses of these types.

22.3.2.1 Sources of Nonindependence

The first issue to address is identifying the sources of nonindependence in your data. You will need to identify each *level of data* as well as whether your data are *nested* or any levels are *crossed*.

22.3.2.1.1 What Do We Mean by Levels?

Often, longitudinal models have two levels of data: repeated measures (seconds, days, years) nested within individuals. The repeated measures are described as being "nested" or "clustered" within individuals and represent the lowest level of data (level 1). The individual represents a higher level (level 2) and is the "clustering" factor. Sometimes, however, your data might have a more complicated structure. If you collected ESM data, you will have multiple measurements each day and multiple days of data. In this case you have three levels of data: momentary reports (level 1) nested within days (level 2) which are nested within individuals (level 3). You might also have more than two levels due to a higher clustering factor. For example, if you are studying teams in an organization and the members of each of your teams complete monthly surveys for a year, then you have three levels of data: monthly surveys (level 1) nested within individuals (level 2) nested within teams (level 3). If you have multiple organizations, you might even have a fourth level (organization).

22.3.2.1.2 What Do We Mean by Nested versus Crossed?

This distinction refers to the structure of the clustering factors. Data are considered *nested* when one factor occurs only within a particular grouping of another factor. For example, in a diary study, a person's daily reports come from that person only. Data are considered *crossed* when observations are nested within multiple clustering factors simultaneously. This can occur with longitudinal data in cases where time itself can be modeled as a clustering factor such that observations collected from different individuals at a specific timepoint are likely to be correlated with each other in some meaningful way. For example, if you are running a study looking at work-related stress and you collect data from participants in the same company for a week, daily reports of stress are nested within individuals. But they can also be nested within the day of the week. Just as an individual's stress level on Monday is likely to be more similar to their level

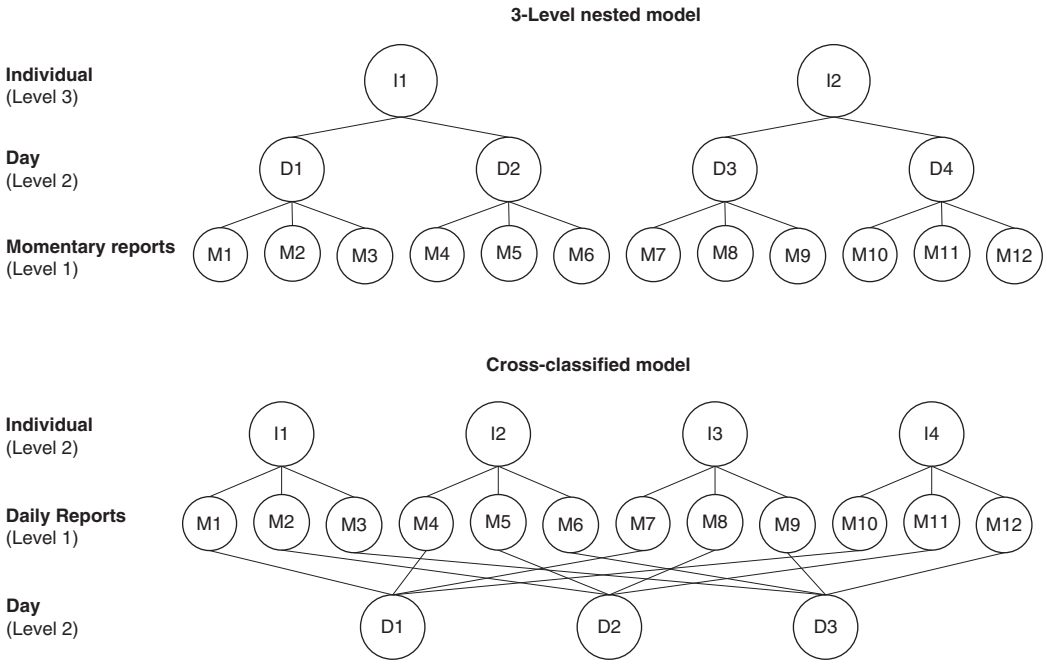

Figure 22.3 Example of nested versus crossed clustering factors

on Saturday compared to someone else's Saturday stress, creating nonindependence within *individuals*, a participant's stress on Monday is likely to be more similar to someone else's stress on Monday than it is to someone else's stress on Saturday, creating nonindependence within *days*. In this model specification, you have two clustering factors: individuals and time, but instead of one factor (e.g., days, level 2) being nested within the other (e.g., individuals, level 3), the two factors are crossed (both individuals and days are at level 2) because the same daily report is nested both within an individual and within a day.

Because crossed factors can be hard to explain in words, Figure 22.3 depicts examples of a nested and a crossed design. Note that the assumption with the three-level nested model in Figure 22.3 is that the momentary assessments occur randomly throughout the day and thus one person's day 1 assessments are unlikely to be related to another individual's day 1 assessments in a meaningful way. Adding some complexity, if they are collected at set times that meaningfully differ from each other (e.g., morning, afternoon, night), *time of day* may also be a clustering factor in this data set (two participants' feelings in the morning may be more similar to each other than one participant's feelings in the morning and the other partner's feelings in the afternoon, creating correlations within time of day).

22.3.3 Different Modeling Options for Nonindependence

You can deal with nonindependence in multiple ways. Often social and personality psychologists use multilevel models (MLMs; also known as mixed-effects models, random-effects models, or hierarchical linear models), which account for the multiple levels and nested or crossed structure of the data. These models take advantage of the unique aspects of repeated-measures

data by modeling fixed effects (i.e., average responses) and random effects (i.e., variability around the average response). Thus MLM is a great tool if you are interested in understanding heterogeneity in effects; with repeated-measures data, people are often interested in understanding whether there is significant variability in how people change over time (e.g., Bolger & Zee, 2019). For example, if you want to explore whether the effects of daily feelings of loneliness on daily mood are similar for everyone in your sample, or whether people vary in how much their mood is affected by feelings of loneliness, MLM allows you to do this. However, sometimes social and personality psychologists are not interested in modeling this heterogeneity and simply want their models to appropriately account for nonindependence (especially when there are many sources of nonindependence and some of them are crossed). If you are interested only in whether feelings of belonging, on average, tend to predict daily mood, and not in whether this association varies from person to person, then there are other statistical approaches you can use to account for nonindependence. These may be less complex than MLM, with fewer assumptions, and require fewer data points because they are less computationally intensive.

22.3.3.1 Alternative Approaches to Dealing with Nonindependence

One approach you can use other than MLM is accounting for nonindependence within clusters through *fixed effects* (i.e., including your clustering factor as a categorical covariate). You may have seen this approach used before, such as when researchers collect global data from a number of different countries (with many participants nested within each country), and then treat country as a categorical covariate. If you have repeated-measures data with only a few timepoints (e.g., two to five) and you think you have cross-classification where time is a clustering factor, you might consider treating time as a categorical covariate. One reason you might take this approach is if you are interested in making comparisons between particular timepoints. For example, if you only have diary data for Monday through Friday, you might be interested in making comparisons between each of the weekdays. If you have ESM data with morning, afternoon, and night assessments and time of day is a source of nonindependence, you may want to model time of day as a categorical covariate. Number of timepoints and interest in comparisons *between* timepoints may go together: when you have a small number of timepoints that differ meaningfully from each other, you may be more likely to have time as a source of nonindependence and to have hypotheses about differences between specific timepoints (see section 22.4 below for information on treating time as a *continuous* covariate).

If you have a small number of timepoints, another approach is to model your data using *repeated-measures ANOVAs*. This approach is fairly easy to implement and can work when you have only a few timepoints; however, it is rarely used anymore. One reason is because it assumes homogeneity of covariances between timepoints (i.e., the correlation between time 1 and time 2 is the same as between time 1 and time 3), which is rarely the case. Repeated-measures ANOVAs also cannot handle missing data (unless you impute values so that data are no longer missing). If you have any data missing from a participant, none of their data will be included (i.e., listwise deletion). Similarly, repeated-measures ANOVAs cannot handle different numbers of timepoints across participants – as might occur, for example, if you measured participant's physiological responses each minute across a task that took participants different amounts of time to complete. Lastly, repeated-measures ANOVAs cannot handle more complicated data structures than a simple two-level design and cannot incorporate time-varying predictors (see Table 22.2). If, for example, you are interested in predicting daily mood from daily feelings of belonging, you must use a different approach. If none of this is a

Table 22.2 Definitions of common terms in over-time, repeated-measures design and analysis

Term	Definition
Between-person effect	Effects that compare one person to another (i.e., a between-person comparison). These result from predictors that have only one score per person (e.g., age, average mood across two weeks)
Clustering or nesting variable	Variable referring to a unit that groups other observations; in repeated-measures data, the clustering variable is usually person and the observations grouped within person are the different timepoints (we also refer to this as a *clustering factor* throughout the chapter)
Cross-classification	Random factors are considered crossed when observations are nested in multiple clustering factors at the same time, such as when measurements are nested within people and within days (see Figure 22.3)
Effective sample size	The effective sample size is the sample size of a simple random sample with independent observations that has the same precision of estimates (and, therefore, statistical power) as a sample with nested data. The higher the ICC, the smaller the effective sample size because additional repeated measures provide less novel information. Effective sample size will be somewhere between the number of repeated measures and the number of units (usually participants, but can be groups, teams, etc.)
Fixed effect	Average effects (intercept, slopes) for your sample
Growth curve model	Statistical model examining change over time
Intraclass correlation (ICC)	The strength of the correlation between observations within a cluster. For longitudinal designs, this typically refers to correlations between repeated measures within a person
Lagged analysis	Predicting one timepoint from a prior timepoint to assess directionality (e.g., predicting mood today from belonging yesterday, controlling for mood yesterday)
Random effect	Variability around an average fixed effect (intercepts, slopes; e.g., if you have repeated measures nested within participants and participant is your random factor, your random effects refer to between-participant variability around the mean levels for participants)
Random factor	In a multilevel model, clustering variables can be specified as random factors, allowing random effects to be estimated. In repeated-measures data, person is often a random factor, which then allows one to estimate person-to-person variation in effects (i.e., random effects for person)
Time-varying variables	Variables that can have different scores at different timepoints; usually, both within-person and between-person effects can be derived from these variables
Time-invariant variables	Variables that do not change across time because they were only assessed once or have the same score at every timepoint; these can only produce between-person effects
Within-person effect	Effects that compare associations from timepoint to timepoint within a person. These often result from predictors that are person-mean-centered, so that each repeated measure from a participant reflects a deviation from their mean score.

problem for you, then this might be an appropriate model given its simplicity.

Another approach you can take is to *directly adjust your residuals* (and thus correct bias in your standard errors) without modeling random effects. This modeling approach parallels typical single-level models, allowing you to interpret the data just as you would if you did not have repeated measures, except that it adjusts for correlated residuals due to nonindependence. "Marginal models," "generalized estimating equations" (GEE), "population-averaged models," and "cluster-robust standard errors" are all terms that refer to this alternative approach to modeling data with nonindependence. Some consider these models to be underutilized in our field – they are appropriate for situations where you need to account for nonindependence but you are not interested in any of the conceptual questions that multilevel or structural equation models can help you answer (for more on the "unnecessary ubiquity of hierarchical linear modeling," see McNeish et al., 2017).

Researchers with longitudinal data may also be interested in using *structural equation models* (SEMs) to capture change over time (McArdle & Nesselroade, 2014; McNeish & Hamaker, 2020). Common SEMs include the latent growth curve model, which looks at change across all timepoints, and cross-lagged models, which look at change from one timepoint to the next within and across variables (see Usami et al., 2020, for a discussion of the best approaches for cross-lagged analyses). SEM can handle nonindependent data because it allows you to specify correlated errors between repeated measures. Why might you choose use SEM over MLM or one of the other approaches described above? Often SEM and MLM yield similar results, leading to personal preference and familiarity with one type of modeling (see Bolger & Laurenceau, 2013 for analyses using both approaches). In general, SEM provides you with more flexibility in specifying your model. For example, in SEM you can easily specify different associations between variables at different timepoints as well as test for measurement invariance across timepoints. It is also easier to develop more complex models with multiple outcomes in SEM. SEM provides more flexibility when comparing models between groups because you can specify associations between variables separately for each group. SEM also does a better job of handling missing data on the predictor side (McNeish & Matta, 2018) and can assess model fit as a whole within a single model, whereas MLM can only compare fit between models (e.g., Ledermann & Kenny, 2017).

There will be times when SEM is *not* the right choice. If you have intensive longitudinal data with many timepoints, an SEM in which you specify every timepoint as a separate observed variable can get unwieldy quickly. If you have complex nonindependence beyond a two-level model, MLM is better set up to handle complicated nesting structures. MLM can also handle random effects, whereas SEM cannot. However, advances in multilevel SEM are making it easier to combine the benefits of both analytic approaches, such as using SEM even with data that have more than two levels or crossed clustering factors. MLM may also be better suited to handling smaller samples.

22.3.3.2 Which Approach Should I Choose?

Which approach you choose will depend on the type of question you are interested in asking. Do you want to say something about heterogeneity between people in how particular processes unfold over time? Or are you only interested in average effects? Are you interested in temporal patterns or just in describing the typical association between two variables? Importantly, you can often make the choice of which approach you use for *each* clustering factor in your model. Let's return to our example of crossed data in which daily reports of stress were nested both within individuals and within days of the week. For days of the week, there are only seven clusters (the seven days of the week). Rather than modeling these as a separate

random factor, you could choose to account for nonindependence within days by taking a fixed-effects approach. You would end up with a two-level MLM in which daily reports are nested within individuals, and weekday is a categorical covariate. With team data, you might choose to model teams as a covariate if you have many participants but they are clustered into a small number of teams. In the case of our teams example above, you would have an MLM with monthly surveys nested within individuals and team would be included as a covariate. This is also the case for SEM: sometimes you might have a complicated data structure that would be difficult to model in SEM, but if you are able to treat some of the clustering factors as fixed effects, then you can simplify your model to two levels. With *dyadic* longitudinal data, researchers often end up adjusting for different aspects of nonindependence using different strategies (for more information on dyadic longitudinal analyses, see Kenny, Ackerman, & Kashy, Chapter 23 in this volume; Kashy & Donnellan, 2008; Chapters 13 and 14 of Kenny et al., 2006; Chapter 8 of Bolger & Laurenceau, 2013; and Thorson et al., 2018).

22.3.4 MLM: Random Factors and Random Effects

Given that MLM is currently the dominant approach for dealing with longitudinal data in social and personality psychology, we focus the remainder of this section on MLM.

First, we have already used the terms "random factor" and "random effect," but let's pause to explicitly discuss what each represents (see Table 22.2 and also Judd & Kenny, Chapter 24 in this volume). We have used the term "clustering factor" or "clustering variable" to describe the structure of repeated-measures data in which variables are nested or crossed. Above, we described several different ways of modeling clustering factors to deal with this nonindependence – for example, using a fixed-effects approach and treating the clustering factor as a categorical covariate in the model. Now, we turn to dealing with nonindependence through MLM, which can account for the nonindependent nature of the data by treating clustering factors as *random factors*. For example, in an MLM analysis using diary data in which days are nested within individuals, the clustering factor – individual – would be treated as a random factor.

Turning to the effects in the model, *fixed effects* refer to the average estimates for your sample, whereas *random effects* refer to variability around that average. What variability? It depends on the random factor. Random effects are specific to a random factor and refer to variability related to that factor. For longitudinal data collected from individual participants, you can estimate random effects at the individual level (level 2; random factor: individual), which models between-person variability in your effects. For example, if you have both fixed and random effects of feelings of belonging on mood, the fixed effects will tell you about how feelings of belonging typically relate to mood in your sample, and the random effects will tell you how much people tend to vary from each other in the association between feelings of belonging and mood. Feelings of belonging may be highly correlated with mood for some individuals, but not for others, and a random effect can capture this between-person variability. In a more complicated MLM, such as one with three levels, you will have multiple random factors and can have random effects for each of these separate factors. For example, if your belonging and mood study included ESM data that was best modeled as timepoints nested within days nested within individuals (three levels), you could address the nonindependence within days and individuals by treating day and individual as two random factors. You could then examine random effects for each random factor, such as whether effects of feelings of belonging on mood varied across days as well as across individuals.

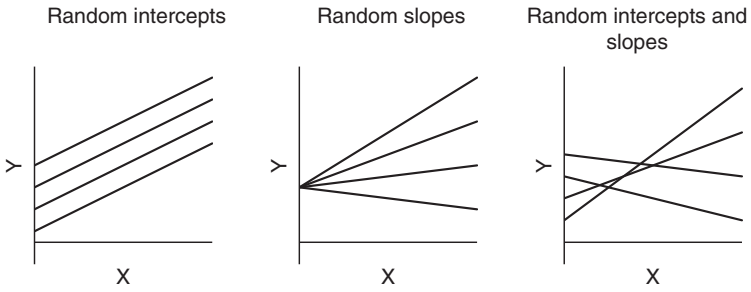

Figure 22.4 Depiction of predicted values from multilevel models with random intercepts, slopes, or both. Each line represents an individual participant

You can choose whether or not to model random effects, but you must have at least one random effect (typically a random intercept) to have a multilevel model. There are three types of random effects (see also Figure 22.4). Below, we explain each type of random effect using participant as the random factor (i.e., random effects for repeated measures nested within participants), but these three types of random effect apply to any random factor.

1 *Random intercept*. When you have a random intercept with repeated measures nested within participants, in addition to modeling the average intercept for the whole sample, your model also estimates intercepts for every person. The variance estimate for the random intercept indicates how much variability in the outcome variable there is in the sample around the fixed intercept. If you model time as a predictor and the intercept represents the first timepoint (see below for more on centering), then if everyone starts at the same level, the random intercept will be small and possibly nonsignificant. If people's outcome scores vary widely at the first timepoint, the random intercept will be large. Thus the size of the random effect can provide information about how generalizable the fixed effect is: does the estimate for the fixed intercept reflect most people or not?
2 *Random slope*. You can only model random slopes for time-varying variables. Time-invariant measures, such as those measured only once per person, cannot have random effects because you cannot calculate individual slopes for each person. For time-varying predictors, you can choose whether or not to allow random effects. If you have a random slope for a predictor, then you are not just modeling the average slope for everyone but also estimating individual slopes for each person. This allows you to capture the variability of those slopes around the average, fixed slope. The estimate for the random slope reflects this between-person variability.
3 *Random covariance*. When you have more than one random effect in your MLM, you can also allow covariances between them. If you have a random intercept and a random slope, modeling the covariance between them will capture any correlation that might exist between different mean levels when the predictor is zero and different slopes (e.g., whether people who start out higher change more). Covariances between two random slopes will capture whether individual variability in one slope is related to individual variability in the other slope (e.g., whether people who show stronger correlations between one predictor and the outcome also show stronger correlations between another predictor and that outcome).

You also have flexibility in specifying whether all covariances between random effects are

modeled or just specific covariances. For example, an unstructured (UN) variance–covariance matrix for random effects will estimate all possible random effects, including all random covariances, whereas a variance components (VC) matrix will estimate all random variances, but specify all random covariances as zero. There are many types of matrices and some programs have default matrices, which are important to understand.

22.3.4.1 Residual Variances and Covariances

In addition to the level 2 random effects specified above, random effects at level 1 (in our example, within-person) can be estimated as well. These impose a structure on the residuals: the differences between the predicted values for a person at a particular timepoint and the actual values for that person at that particular timepoint. There is a variance–covariance matrix for these residuals, and you can make decisions about the structure of estimations in this matrix. To begin, you can make decisions about the *variances*. For example, you can specify whether the residual variances (i.e., error variance) should be estimated as equal (estimating only one variance such that every timepoint has the same error variance) or heterogeneous (estimating one per timepoint; e.g., fourteen different variances, one for each day, in a two-week diary study). If you have many data points, heterogeneous variances can be computationally intensive and reduce dfs. You can use model testing to make decisions about your best-fitting model.

You can also specify certain structures for the *covariances* of these residuals. With repeated-measures data, there are almost always time-related correlations in residuals. For example, if participants report on negative affect daily, it is likely that days which are closer together will show more highly correlated measurements than those that are farther apart – this is known as autocorrelation. If this temporal patterning is not captured elsewhere in your model (e.g., with lagged fixed effects), then you need to consider modeling these patterns in the residuals. Because these correlations can be quite powerful in repeated-measures data, if you ignore them you risk biased standard errors for your fixed effects, which can produce Type I errors (Greene, 2008). The most common error structure for repeated-measures data is a *first-order autoregressive structure* in which the variance of errors across timepoints is the same (i.e., the same error variance at each timepoint, as described above), and covariances between errors with the same time lag are the same (i.e., T1 to T2 is the same as T3 to T4, and T1 to T3 is the same as T3 to T5; Wu et al., 2013). There are a few other variance–covariance structures that may be relevant to over-time data, such as a Markov structure for unequally spaced data. Potential relationships between residuals over time is one reason why it is important to be thoughtful about how you structure your time variables, especially when they might have unequal spacing.

An example. Let's return to our belonging and mood example with daily data collected across a week. You have a random intercept and a random slope for belonging. As in a typical regression model, the *fixed* intercept tells you, on average, what people's moods are when their feelings of belonging are 0 (see below for more on centering), and whether that average level is significantly different from 0. The *random* intercept tells you whether people's moods vary when belonging is 0: does everyone feel the same or do people have differing moods when their feelings of belonging are 0? The *fixed* effect for belonging tells you the average association between belonging and mood. Does mood tend to become more positive as feelings of belonging increase and does this slope differ significantly from 0? The *random* effect (i.e., random slope) for belonging tells you how much individuals vary in terms of how their feelings of belonging influence their mood. Does belonging

affect everyone's mood the same way, or does it have stronger effects on some people than on others? A *random covariance* between the random intercept and the random slope for belonging tells you whether people who are higher or lower in terms of their mood when belonging is 0 tend to show stronger or weaker relationships between belonging and mood.

Let's also add time as a predictor to see whether people's moods change meaningfully over the week. If we have a random effect for time, then the fixed and random intercepts now *also* represent average mood when time is 0 and between-person variability in mood when time is 0 (i.e., when time and belonging are *both* 0). The fixed and random slopes for time represent the average change in mood across time as well as the individual variability in this change: How does mood tend to change over the week? And does mood change differently for different people? The random covariance between the random intercept and the random slope for time represents the associations between between-person variability in mood when time and belonging are 0 (random intercept) and the variability in changes in mood across time (random slope for time). There can also be a random covariance between belonging and time if there are random slopes for both predictors. This random covariance indicates whether the individual associations between belonging and mood are correlated with the individual associations between time and mood – that is, do people who have a stronger relationship between belonging and mood experience stronger or weaker changes in mood over time compared to those with a weaker relationship between belonging and mood? Lastly, in order to help explain any lingering error in our data, we can impose a structure on the level 1 residuals – specifying that residuals on adjacent days are more highly correlated than those on days further apart (an autoregressive structure). We can also impose one residual variance across all timepoints, which indicates whether there is any remaining variance in the residuals that has not yet been accounted for (as noted above, we could also have heterogeneous variances that are independently predicted for each timepoint, but that tends to be a computationally cumbersome model if there are many timepoints).

Table 22.3 Terms in Equation 22.1

Term	Description	
Y_{ij}	Outcome (in this example, mood) for time i for person j	What is person j's mood at time i?
β_{0j}	Intercept for person j	What is the average mood for person j?
β_{1j}	Slope for belonging for person j	What is the average relationship between belonging and mood for person j? (I.e., how much does person j's mood change when they feel they belong more?)
X_{1ij}	Belonging for person j at time i	What is person j's belonging at time i?
B_{2j}	Slope for time for person j	What is the average relationship between time and mood for person j? (I.e., how much does person j's mood change over time?)
X_{2ij}	Time for person j at time i	What is the value for time for person j at time i?
r_{ij}	Residual or error at time i for person j	What is the difference between actual mood at time i for person j and model-predicted mood at time i for person j?

Although we have seen many people fear equations, they can be a very useful tool for understanding multilevel models and explaining them to readers. We use the example above to introduce readers to the Raudenbush & Bryk (2002) notation, given its frequency within psychology.

The level 1 or within-person equation for the example model we described above is as follows:

$$Y_{ij} = \beta_0 j + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + r_{ij} \qquad (22.1)$$

Parts of the level 1 equation can be broken into multiple components, as shown in these level 2 equations (Equations 22.2–22.4). These equations make clear that the intercept and both of the slopes (for belonging and time), for any given person in the study are combinations of fixed and random effects.

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad (22.2)$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \qquad (22.3)$$

$$\beta_{2j} = \gamma_{20} + u_{2j} \qquad (22.4)$$

Inserting the level 2 equations into the appropriate locations in the level 1 equation yields one equation for the whole model (Equation 22.5):

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{1ij} + \gamma_{20} X_{2ij} + u_{0j} + u_{1j} X_{1ij} + u_{2j} X_{2ij} + r_{ij} \qquad (22.5)$$

One aspect of the model that is not apparent from the equations above is the specification for the variance–covariance matrix of both the level 2 random effects and the residuals. Recall that, when reporting these, you are unlikely to report "subject-specific" random effects – meaning the specific deviations for a particular effect for each participant (e.g., u_{0j} *or* u_{1j} *or* u_{2j}). Instead, you would indicate the variances and covariances of these effects because these tell readers whether there is significant deviation from the fixed effects overall (variances) or whether the deviation for one effect tends to be related to the deviation for another effect (covariances). We display notation for these in Table 22.5.

Table 22.4 New terms (i.e., not used in Equation 22.1) in Equations 22.2–22.4

Term	Description	Estimated as fixed or random
γ_{00}	Intercept	Fixed
u_{0j}	Deviation in the intercept for person *j*	Random
γ_{10}	Slope for belonging	Fixed
u_{1j}	Deviation in the slope for belonging for person *j*	Random
γ_{20}	Slope for time	Fixed
u_{2j}	Deviation in the slope for time for person *j*	Random

Table 22.5 Notation for the level 2 random effects

Term	Description
τ_{00}	Variance in the intercepts
τ_{11}	Variance in the belonging slope
τ_{22}	Variance in the time slope
τ_{01} or τ_{10}	Within-person covariance between each person's intercept and their belonging slope
τ_{02} or τ_{20}	Within-person covariance between each person's intercept and their time slope
τ_{12} or τ_{21}	Covariance between one person's belonging slope and their time slope

How do you know whether to include a random effect in the model? The conservative approach is to start with a fully unstructured random variance–covariance matrix in which you model all random intercepts, slopes, and covariances (as shown in the example above in Table 22.4). Sometimes these models have difficulty running due to complexity or lack of variability in a random effect. In this case, you may need to remove problematic random effects (see Brauer & Curtin, 2018 for guidance on removing random effects) or turn to a Bayesian approach, which can better

handle this complexity. If you are interested in trimming random effects, many scholars recommend against conventional significance tests for random effects due to multiple issues with estimating significance (distribution, power, and so on; Barr et al., 2013; Nezlek, 2012; Thorson et al., 2018). The final model you use will be guided by your conceptual questions and the practicalities of which effects you can robustly estimate. We recommend reporting which random effects you estimated (including variances and covariances) and explaining your decision making, especially if you chose not to include certain ones.

22.3.5 MLM: Centering

With longitudinal data, there are multiple ways to center your variables, and these different approaches yield meaningfully different results. With your time variable, it is especially important to make sure that zero is meaningful. Often people choose to center on the first timepoint, so that the intercept represents people's starting point. But you can also center on the midpoint, end, or some other meaningful timepoint. You can also choose whether zero means the same thing for everyone in your sample. For example, in a diary study where people started on different days of the week, zero could be everyone's first day, or the first Sunday for everyone. The key is to be thoughtful about what zero represents, as this will influence how you interpret your fixed and random effects. For instance, the random intercept reflects variability when predictors are zero. Thus you can get strange estimates of random effects if zero is not represented in your data, as is often the case when we measure variables using Likert scales that begin at 1. For more on centering, see Bolger & Laurenceau (2013); Enders & Tofighi (2007); Hamaker & Muthén (2020).

22.3.5.1 Unconfounding Within- and Between-Person Effects

Longitudinal data confound within- and between-person effects. Here we explain exactly what that means. When thinking about the effects of belonging on mood, if people are in a more positive mood when they feel greater belonging across a week, it might be due to the fact that people who tend to feel they belong more than the average person experience more positive moods than people who tend to feel they belong less than the average person. This is a *between-person effect*, comparing associations from one person to another. It might also be that on days when people feel they belong more than they usually do, they are in a more positive mood than on days when they feel they belong less than they usually do. This is a *within-person effect*, comparing associations from timepoint to timepoint within a person. Without properly centering the data, you cannot tell whether the fixed effects in your MLM results are due to between-person effects, within-person effects, or both.

One way you can unconfound within- and between-person effects is by *person-centering* your data. You do this by getting each person's average score for a predictor and subtracting it from each of their repeated measures to create a within-person centered variable; this variable captures fluctuations relative to each person's average. The variable containing the average score for each person then represents the between-person effect (usually this gets grand-mean-centered so that zero is meaningful). Entering these two variables as predictors separates within- and between-person effects. You can also look at the interaction between them to see whether people who are higher or lower on average tend to show different associations at the within-person level. For example, some research shows that people who tend to experience more relatedness in their lives experience greater boosts in mood on days when their experiences of relatedness are greater than usual, relative to people who generally experience less relatedness in their lives (Moller et al., 2010).

Separating within- and between-person effects can be conceptually informative. You

may find that some effects are stronger at one level than the other, and in rare cases they can even be reversed. For example, generally feeling inauthentic may be worse for well-being than feeling less authentic on a single day. In contrast, people may be in a worse mood after sleeping less than they usually do, but shorter sleepers may not typically be in a worse mood than people who tend to sleep longer. Lastly, if you are interested in knowing whether the within-person and between-person effects differ significantly from each other, you can test that using a *contextual* model (see Hamaker & Muthén, 2020, for details).

22.3.6 Moderation and Mediation

22.3.6.1 Moderation

Because you can separate data into within- and between-person effects, you can also run moderations at each of these levels, as well as across levels. *Cross-level moderation* occurs when you look at whether a within-person effect varies as a function of some between-person variable, such as our example above of whether the daily within-person effects of belonging on mood are different for people who tend to have higher or lower levels of belonging. You can also have cross-level moderation with two different variables. For example, you might want to see whether a person's social status moderates the within-person association between belonging and mood.

Within-person moderation looks at interactions between two time-varying variables. Perhaps on days when people experience more academic stress than they typically do, feelings of belonging have a stronger association with mood than on days when they experience less academic stress than usual. For these within-person moderations, you have to keep in mind that you can model random effects for each predictor, as well as their interaction. Essentially, this allows the interaction between these two variables to vary from person to person.

You can also have *between-person moderation*, which will have no random effects. For example, you might be interested in testing whether people who tend to have lower levels of belonging than others (i.e., between-person differences in average belonging) are buffered from a more negative mood if they have higher social status (another between-person variable).

Be aware that if you do not unconfound within- and between-person effects for time-varying variables by creating variables that represent each person's average and their variability around their average (i.e., person-centered), then moderation results will also be confounded.

22.3.6.2 Mediation

Unconfounding within- and between-person effects can also be important with longitudinal mediation to ensure that your mechanism is specified at the same level as, or a lower level than, your predictor. For example, self-esteem, a between-person (level 2) variable, cannot explain daily changes in mood (level 1) as a function of daily fluctuations in belonging (level 1). Instead, any mediator of within-person changes in mood must also be within-person. Thus you typically want to separate out within- and between-person effects and conduct mediation analyses at the appropriate level with the correctly centered variables.

If you are conducting mediations that include both a predictor and a mediator that are within-person, then you can model random effects for the association between the predictor and the mediator (i.e., the 'a' path) and between the mediator and outcome (i.e., the 'b' path; see Montoya, Chapter 25 in this volume). If you model these as two random effects, then you allow the two mediational paths, a and b, to vary from person to person. In this case, even when there is an indirect effect on average, a substantial proportion of participants may not show an indirect effect,

suggesting more limited within-person mediation. For example, perhaps people feel more socially secure on days when they feel more belonging, and this lifts their mood. We might see this indirect effect when we look at the fixed effects, but the random effects could tell a different story. For one person, it may be that belonging is associated with more security, but security is not associated with mood. For another person, belonging is not associated with feeling more secure, but security does predict increased positive mood. In order to show within-person mediation when you are modeling random effects, you therefore need to run an analysis in which you simultaneously predict both your indirect paths and the covariance between their random effects (see Bauer et al., 2006 for steps for running this type of analysis).

22.3.7 Deciding How to Structure Your Time Variables

If you have equally spaced timepoints, creating a variable to represent time is simple (e.g., 0–6 for a seven-day diary). However, if you have unequal spacing, you have to decide whether to model the timepoints by the number of the timepoint (0, 1, 2, 3) or distance from baseline (0, 1, 6, 12). Which approach you take will depend on the questions you are asking, but be aware that it is important to model time as actual distance from baseline in situations where you are (a) accounting for autocorrelations that assume that errors in timepoints closer together are more highly correlated than those further apart and/or (b) modeling time as a predictor, such as when using growth curve models.

When you have event-contingent data, you will have unequal spacing, with data at different times for different people. You could average across time bins (e.g., if you track people for a month, you could calculate the average events for each person for each of the four weeks). Or you can treat time as an ordering variable, listing each event sequentially, and then create a second variable that lists when each event occurred for each person, relative to a point of interest, such as the baseline or the time the event last occurred (e.g., how many minutes since a person last checked social media). Sometimes you have different numbers of events per person and need to create a variable for time that reflects the largest number of events possible. For example, if people report every time they have a conflict, your time variable would go from 1 to 20 (i.e., twenty rows or columns) if the maximum number of conflicts reported in your sample was twenty, and most people would have some missing data.

You can create multiple time variables to reflect different ways of modeling time. If you have ESM data, you might have one variable that represents data collected within a day (e.g., 0–3 for the four daily check-ins), one that represents each day (e.g., 0–6 for the seven days), and one that is sequential (0–27 for all timepoints). These different approaches provide you with flexibility in your analyses.

22.4 Additional Considerations with Repeated-Measures Data

When modeling over-time data, there are advantages and challenges. Below we outline some of the unique questions you can ask with repeated-measures data, as well as providing details on additional analytic issues.

22.4.1 Assessing Over-Time Patterns

As we have mentioned throughout this chapter, longitudinal designs that have many timepoints create opportunities for understanding how processes change over time. For example, relationships researchers are often interested in mapping patterns of long-term change in relationship quality: work on newlyweds, for instance, has examined how personality traits and behaviors relate to change in marital satisfaction over time (Karney

& Bradbury, 1997; Lavner & Bradbury, 2010; Williamson & Lavner, 2020) The most common approaches for modeling change over time include growth curve models in which time is entered as a predictor in the model. There are both MLM and SEM-based growth models (repeated measures are modeled as indicators on a latent factor in SEM). Although growth curve models may sound complicated, they are actually fairly straightforward. Time is a predictor in your model, and you can examine mean levels (via the intercept) as well as change over time (via the slope for your time variable). You can also have moderators, to test whether changes over time differ as a function of another variable. For example, researchers could choose to test whether change in relationship satisfaction over time depends on attachment security, with securely attached individuals showing more stable patterns of satisfaction. This can be tested with a cross-level interaction (attachment by time). Growth curve models can also accommodate interactions with time-varying variables. Here, it may be easier to think about time as the moderator. For example, researchers could test whether the association between physical attraction and relationship satisfaction changes over time – do people's feelings about their relationship become less tied to how attractively they view their partners as the relationship progresses? Models of these types can be used with more intensive longitudinal data as well, such as diary or ESM data (Bolger & Laurenceau, 2013).

Over-time patterns may not be linear, so it is important to look beyond linear to polynomial or nonlinear associations (Girme, 2020; Hayes et al., 2007). Keep in mind that you must have enough timepoints to fit more complicated polynomial relationships. When visually inspecting your data, you may notice nonlinear longitudinal trends that necessitate the use of non-polynomial (e.g., quadratic or cubic) terms. In this situation, to best approximate different trends in your data, you could use a piecewise regression model, also called a spline, segmented, or broken-stick regression, in which you estimate different slopes for different phases (e.g., Frost & Forrester, 2013; see Simonsohn, 2018, for an algorithm to help identify these slopes for U-shaped trajectories).

When running over-time models, all our points about random effects and centering still stand. In fact, people are often interested in random effects when modeling over-time patterns, because they provide useful information about how much variability there is in change over time (Bolger & Zee, 2019). One somewhat unique feature of time is that, although you have a different score at each timepoint and you can model random effects such that people have different slopes over time, you generally cannot unconfound within- and between-person effects, because everyone typically has the same scores on time at each timepoint (i.e., 0–13 for a fourteen-day diary), and thus they have the same between-person means and deviations. Because of this, time is typically centered on a timepoint of interest (e.g., starting point, midpoint, or point of intervention).

22.4.2 Including Time in Your Model

You may have longitudinal data in which time is not a central interest, such as our prior examples of predicting mood from feelings of belonging. Although you have over-time data, your research questions are not actually time-based. Even when this is the case, it is important to consider time as a potential covariate. For our example, it may be that as people complete daily diaries reflecting on feelings of belonging and their mood, they become more sensitive to their environment and their mood and this leads to increases in both belonging and mood. If this was the case, then time would be a third variable associated with similar changes in both feelings of belonging and mood. An analysis without time in the model would show a strong association between belonging and mood, but adjusting for time as a continuous covariate would reveal that it was a

spurious correlation. Treating time as a covariate accounts for direct effects of time on these variables, as well as any other relevant processes that also change systematically over time. Note that when time is treated as a continuous covariate, it accounts for over-time trends in the outcome, but does not account for any nonindependence *within* or *between* timepoints, as described in the sections above. In models where you have nonindependence within timepoints or have correlated residuals, you can both model time as a predictor of interest and account for its nonindependence.

As described above, in addition to treating time as a main effect in your model, you can consider interactions between time and other predictors (i.e., growth curve analyses). For example, researchers found that the within-person association between co-rumination and rumination increased for adolescents over the course of three and a half years (DiGiovanni et al., 2022). Temporal changes in within-person associations may also be the result of measurement error: for instance, researchers found that the association between a child's reported mood and reported conflict with their parents weakened over time, an effect they attributed to fatigue with the study (Reynolds et al., 2016). Again, temporal change in predictor–outcome associations might not be your key question of interest, but understanding whether it exists may be important for thoroughly characterizing your effects.

22.4.3 Assessing Directionality

When you collect repeated measures, one opportunity you have is testing questions of directionality using lagged models (though note that you can also assess questions of directionality with repeated-measures data using simultaneous-effects models; Goldring & Bolger, 2021). Although lagged models certainly do not replace experimental manipulations, they can help inform understandings of directional relationships (Iida et al., 2012). Lagged models predict a variable at one timepoint from a variable at a prior timepoint. For example, to identify the direction of the relationship between feelings of belonging and mood, you can predict mood tomorrow from feelings of belonging today and vice versa (feelings of belonging tomorrow from mood today). If you find that feelings of belonging today predict mood tomorrow but not the reverse, then you have evidence for a direction that goes from belonging to mood (this does not mean that there is strong evidence that belonging causes mood, but rather that there is a temporal order to these processes and that we can predict some of the variability in a person's mood on one day from their feelings of belonging the prior day). If you find evidence for both directions, that might suggest a bidirectional association. Importantly, these models should adjust for the outcome variable measured at the same time as the predictor, in order to predict change over time and adjust for correlations between the outcome measured at different timepoints. Thus, in a lagged model with belonging predicting mood, you would have belonging today and mood today predicting mood tomorrow. You can do this in models with only two data points, or with many points, in which every timepoint predicts the next timepoint.

The examples above describe models in which today's predictor predicts tomorrow's outcome, but you may also read articles in which yesterday's predictor predicts today's outcome – these are the same model. You can also predict lags of greater than one day. Perhaps you are interested in the lingering effects of belonging. You can test this by conducting analyses in which mood is predicted from belonging one, two, or three days earlier (e.g., Bolger et al., 2000). Be aware, however, that you will "lose" data as your lags get longer: a lag of one is associated with one lost timepoint, a lag of two is associated with two lost timepoints, and so on. This data loss is also compounded by any missing data. For instance, with a lag of two timepoints, every missing timepoint deletes two timepoints; thus, a fourteen-day daily

diary study with a lag of two and two missing timepoints on days 5 and 8 has only eight timepoints for which both predictor and outcome values exist.

Another approach that provides some evidence of directionality, though not as strongly, is to run models in which your predictor and outcome *are measured at the same timepoint*, but you adjust for the outcome at the prior timepoint. In this way, you still assess change over time, but look at the associations between your variables of interest within the same time period. For example, if you are interested in associations between belonging and mood within the same day, adjusting for mood the prior day can help you test whether it is just that people feel better, both in terms of belonging and mood, following days when they were in a better mood, or whether belonging is uniquely associated with mood above and beyond its prior effects. In other words, on days when you feel you belong more than you usually do, do you experience increases in mood from the prior day? In this way, you can glean some evidence that points to directionality even when your interest is in contemporaneous associations.

A third way people model lagged effects is looking at change in one variable predicting change in the other variable (see, e.g., Stadler et al., 2012). For our diary study with belonging and mood, this would look like a change score from today's feeling of belonging to tomorrow's feeling of belonging predicting a parallel change score from today's mood to tomorrow's mood. Note that you will get different effects if you also adjust for mean levels of belonging and mood today because change may be related to initial levels (see section 22.4.6 on pre–post design below for more on this). This approach provides information about whether corresponding changes are associated with each other, providing slightly different information about causality than the other lagged approaches.

The best approach for examining directionality will depend on the question you are interested in testing. It is appropriate to test lagged effects in multiple ways, as long as you are transparent in your reporting (e.g., Matthews et al., 2014; Orth et al., 2021). In fact, conducting lagged analyses of different types may yield valuable insights into the ways in which your variables of interest are (or are not) related to each other over time. As with other aspects of longitudinal data analysis (e.g., random effects estimation and nonlinear trends), preregistering your exact analysis plans for lagged effects might be difficult. Therefore, if you want to preregister your analysis strategy, we recommend preregistering a general plan with specific steps that build on each other. For example, maybe you plan to examine temporal variability in your data first and then to test a certain lag length based on what you learn. Your preregistration could also include the kinds of lagged analysis you plan to test and a statement that you will report all analyses conducted.

22.4.4 Data Visualization

As with all kinds of data and models, visualizations can be powerful tools. With repeated-measures data, we recommend visualizing your data at several steps in your analytic process. First, examining visuals of one's raw data can be incredibly useful. Histograms of the variables you've measured, as well as raw plots of how variables change over time, can help you understand the nature of the processes you're examining. Many times, these plots can inform choices within your analytic model. For example, you might predict that stress increases linearly over the course of the semester for college students. But when you examine the average pattern of change over time, you instead see a nonlinear pattern, leading you to examine time in both linear and nonlinear forms. As another example, perhaps you think that most people experience similar changes in cortisol concentrations over

the course of the day. But, when you examine plots of cortisol levels for individual participants, you notice substantial variability. This observation might lead you to explicitly allow for heterogeneity in people's cortisol responses across the course of a day within your model.

Second, once you begin constructing and evaluating analytic models, it can be helpful to visually assess the degree of alignment between raw values and model-predicted values. These visualizations can provide you with a sense of how well your model fits your data and whether there are any potential issues to look out for. For example, are there a few outlying observations in your data that are driving effects (McClelland, 2000)? Does your model appear to be a good fit for some participants, but not for others? How big are your effects – they may be statistically significant, but are they noticeable when looking at your data? Many software programs also have particular procedures or packages that provide visualizations aimed at helping you understand whether your model fits your data well and/or whether you have violated assumptions of your model, and these procedures can save an enormous amount of time – and potential embarrassment at a later stage – when evaluating your analyses.

Lastly, there is nothing quite so helpful to other people who want to understand your data as a well-constructed visualization. Many times, visuals are what stick in people's minds after they have read (or quickly scanned) a paper, and so it can be worthwhile to invest time and effort in the development of compelling figures. With repeated-measures data, in particular, readers often want to see how variables change over time (even if this is not a focus of the paper) – both in terms of the average change over time, and in how much variability exists around this effect. Increasingly, readers also expect plots that show predicted values as well as raw observations, and there are many creative ways to show both within the same figure.

22.4.5 Extracting Individual Slopes

If you are using MLM, you can model individual variability through random slopes in your model, and you can also extract those individual slopes and use them as predictors. For example, stress researchers are sometimes interested in whether differences in how people feel on days with high stress versus low stress predict long-term health. In order to test this question, they use daily stress (often coded as stress day versus non-stress day) to predict an outcome of interest, such as blood pressure, and model random slopes. They then extract these individual slopes and use them to predict long-term health outcomes, testing whether people who show greater stress reactivity have worse health over time (e.g., Sin et al., 2015). Another example would be extracting people's individual growth curves to see whether, for example, individual trajectories of satisfaction over the first ten weeks of a relationship predict breakups months later (Arriaga, 2001). Slope extraction is usually an option you can request when setting up the syntax for your model.

22.4.6 Pre–post Designs and Lord's Paradox

When you are interested in predicting change across two timepoints for two or more groups of people, as in a pre–post design, you can use the change between T1 and T2 as the outcome variable (known as a change score approach) or just the value at T2 as the outcome variable. With this second approach, when predicting T2, people often include T1 as a predictor in an attempt to adjust for its influence; this is known as a residualized approach. When you adjust for T1 as a predictor (regardless of whether the outcome is a change score or just the value at T2 – you will get the same results for both of these analyses if T1 is a predictor), the results can be different from an approach where you do not adjust for T1. When comparing groups, this phenomenon is known as

Lord's paradox and can make it challenging to figure out which approach to use. A key question to consider is whether the groups are "pre-existing" and were already different on your variable of interest at T1. For example, if you want to know whether people from rural versus urban areas change over time in the number of daily cross-race interactions they have, you will likely already have pre-existing group differences, with people from urban areas having more cross-race interactions at T1. Assuming that these differences are consistent at T2, you are best off using the change score approach that subtracts group differences out of the outcome (Van Breukelen, 2013). This is because the residualized change approach cannot be used to appropriately adjust for differences that are constant over time between groups. If there are no differences between your groups at T1 (as would be expected with randomly formed groups), then both methods will produce the same results (Van Breukelen, 2013).

22.5 Concluding Thoughts

Although we have tried to be comprehensive in covering the most critical issues of longitudinal design and data analysis in this chapter, no single chapter can provide you with all the information you need to analyze repeated-measures data, nor can it cover all the different possibilities available to you with this type of research design. Thus, below, we list a few additional methods that we have not covered here, along with some recommended books. We also note that this is an emerging area with frequent advances in statistical techniques and analytic programs, which means that there are increasingly new and exciting opportunities that we were not able to cover here. We hope, however, that this chapter will be a resource you can return to as you work to figure out which questions you need to answer when designing and analyzing longitudinal studies.

Additional methods that were not discussed here include

- dynamic structural equation models (Asparaouhov et al., 2018; McNeish & Hamaker, 2020)
- differential-equation/dynamical-systems models (Boker, 2012; Butler & Barnard, 2019; Zee & Bolger, 2022),
- survival analysis (Singer & Willett, 2003),
- time series analyses (Box et al., 2015),
- Markov transition models (Liang et al., 2021), and
- group-iterative multiple-model estimation (Gates & Molenaar, 2012).

22.5.1 Recommended Books

Bolger, N., and Laurenceau, J.-P. (2013). *Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research*. Guilford Press.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied Longitudinal Analysis*. Wiley.

Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. Guilford Press.

Mehl, M. R., and Conner, T. S. (eds.) (2012). *Handbook of Research Methods* for *Studying Daily Life*. Guilford Press.

O'Connell, A. A., McCoach, D. B., and Bell, B. A. (eds.). (2022). *Multilevel Modeling Methods* with *Introductory and Advanced Applications*. Information Age Publishing.

References

Arend, M. G., and Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods*, 24, 1–19.

Arriaga, X. B. (2001). The ups and downs of dating: Fluctuations in satisfaction in newly formed romantic relationships. *Journal of Personality and Social Psychology*, 80(5), 754–765.

Asparouhov, T., Hamaker, E. L., and Muthén, B. (2018). Dynamic structural equation models.

Structural Equation Modeling: A Multidisciplinary Journal, 25(3), 359–388.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68 (3), 255–278.

Bauer, D. J., Preacher, K. J., and Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, 11(2), 142–163.

Boker, S. M. (2012). Dynamical systems and differential equation models of change. In *APA Handbook* of *Research Methods* in *Psychology*, vol. 3, *Data Analysis and Research Publication*American Psychological Association.

Bolger, N., and Laurenceau, J.-P. (2013). *Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research*. Guilford Press.

Bolger, N., and Shrout, P. E. (2007). Accounting for statistical dependency in longitudinal data on dyads. In T. D. Little, J. A. Bovaird, and N. A. Card (eds.) *Modeling Contextual Effects in Longitudinal Studies*. Lawrence Erlbaum Associates Publishers.

Bolger, N., Stadler, G., and Laurenceau, J.-P. (2012). Power analysis for intensive longitudinal studies. In M. R. Mehl and T. S. Conner (eds.) *Handbook of Research Methods for Studying Daily Life*. Guilford Press.

Bolger, N., and Zee, K. S. (2019). Heterogeneity in temporal processes: Implications for theories in health psychology. *Applied Psychology: Health and Well-Being*, 11(2), 198–201.

Bolger, N., Zuckerman, A., and Kessler, R. C. (2000). Invisible support and adjustment to stress. *Journal of Personality and Social Psychology*, 79(6), 953–961.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

Brock, R. L., and Lawrence, E. (2008). A longitudinal investigation of stress spillover in marriage: Does spousal support adequacy buffer the effects? *Journal of Family Psychology*, 22(1), 11–20.

Butler, E. A., and Barnard, K. J. (2019). Quantifying interpersonal dynamics for studying socio-emotional processes and adverse health behaviors. *Psychosomatic Medicine*, 81(8), 749–758.

Chun, C. A. (2016). The expression of posttraumatic stress symptoms in daily life: A review of experience sampling methodology and daily diary studies. *Journal of Psychopathology and Behavioral Assessment*, 38(3), 406–420.

DiGiovanni, A. M., Fagle, T., Vannucci, A., Ohannessian, C. M., and Bolger, N. (2022). Within-person changes in co-rumination and rumination in adolescence: Examining heterogeneity and the moderating roles of gender and time. *Journal of Youth and Adolescence*, 51 (10), 1958–1969.

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F., and Schabenberger, O. (2008). An R2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine*, 27(29), 6137–6157.

Enders, C. K., and Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12, 121–138.

Foster, K. T., and Beltz, A. M. (2022). Heterogeneity in affective complexity among men and women. *Emotion*, 22(8), 1815–1827.

Fox, J. (2015). *Applied Regression Analysis and Generalized Linear Models*. Sage.

Frost, D. M., and Forrester, C. (2013). Closeness discrepancies in romantic relationships: Implications for relational well-being, stability, and mental health. *Personality and Social Psychology Bulletin*, 39(4), 456–469.

Gable, S. L., and Reis, H. T. (1999). Now and then, them and us, this and that: Studying relationships across time, partner, context, and person. *Personal Relationships*, 6(4), 415–432.

Garson, G. D. (2019). *Multilevel Modeling*. SAGE Publications, Inc.

Gates, K. M., and Molenaar, P. C. M. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, 63(1), 310–319.

Girme, Y. U. (2020). Step out of line: Modeling nonlinear effects and dynamics in close-relationships research. *Current Directions in Psychological Science*, 29(4), 351–357.

Goldring, M. R., and Bolger, N. (2021). Physical effects of daily stressors are psychologically mediated, heterogeneous, and bidirectional. *Journal of Personality and Social Psychology*, 121, 722–746.

Gordon, A. M. (2023). Within-person variance in daily conflict and relationship satisfaction. Unpublished data.

Gordon, A. M., and Chen, S. (2014). The role of sleep in interpersonal conflict: Do sleepless nights mean worse fights? *Social Psychological and Personality Science*, 5, 168–175.

Gordon, A. M., Cross, E., Ascigil, E., Balzarini, R., Luerssen, A., and Muise, A. (2022). Feeling appreciated buffers against the negative effects of unequal division of household labor on relationship satisfaction. *Psychological Science*, 33(8), 1313–1327.

Greene, W. H. (2008). *Econometric Analysis*, 6th ed. Pearson/Prentice Hall.

Hamaker, E. L., and Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, 25, 365–379.

Harris, P. E., Gordon, A. M., Dover, T. L., Small, P. A., Collins, N. L., and Major, B. (2022). Sleep, emotions, and sense of belonging: A daily experience study. *Affective Science*, 3(3), DOI:10.1007/s42761-021-00088-0.

Hayes, A. M., Laurenceau, J.-P., Feldman, G., Strauss, J. L., and Cardaciotto, L. (2007). Change is not always linear: The study of nonlinear and discontinuous patterns of change in psychotherapy. *Clinical Psychology Review*, 27 (6), 715–723.

Hox, J., Moerbeek, M., and van de Schoot, R. (2018). *Multilevel Analysis: Techniques and Applications*, 3rd ed. Routledge.

Iida, M., Shrout, P. E., Laurenceau, J.-P., and Bolger, N. (2012). Using diary methods in psychological research. In H. Cooper et al. (eds.) *APA Handbook of Research Methods In Psychology*, vol. 1, *Foundations, Planning, Measures, and Psychometrics*. American Psychological Association.

Karney, B. R., and Bradbury, T. N. (1997). Neuroticism, marital interaction, and the trajectory of marital satisfaction. *Journal of Personality and Social Psychology*, 72(5), 1075–1092.

Kashdan, T., and Steger, M. F. (2006). Expanding the topography of social anxiety: An experience-sampling assessment of positive emotions, positive events, and emotion suppression. *Psychological Science*, 17(2), 120–128.

Kashy, D. A., and Donnellan, M. B. (2008). Comparing MLM and SEM approaches to analyzing developmental dyadic data: Growth curve models of hostility in families. In N. A. Card, J. P. Selig, and T. D. Little (eds.) *Modeling Dyadic and Interdependent Data in the Developmental and Behavioral Sciences*. Routledge.

Kenny, D. A., Kashy, D. A., and Bolger, N. (1998). Data analysis in social psychology. In D. T. Gilbert, S. T. Fiske, and G. Lindzey (eds.) *The Handbook of Social Psychology*, vol. 1. Oxford University Press.

Kenny, D. A., Kashy, D. A., and Cook, W. L. (2006). *Dyadic Data Analysis*. Guilford Press.

Killip, S., Mahfoud, Z., and Pearce, K. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *Annals of Family Medicine*, 2(3), 204–208.

Lafit, G., Adolf, J. K., Dejonckheere, E., Myin-Germeys, I., Viechtbauer, W., and Ceulemans, E. (2021). Selection of the number of participants in intensive longitudinal studies: A user-friendly shiny app and tutorial for performing power analysis in multilevel regression models that account for temporal dependencies. *Advances in Methods and Practices in Psychological Science*, 4(1), https://doi.org/10.1177/2515245920978738.

Lane, S. P., and Hennes, E. P. (2018). Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships*, 35(1), 7–31.

Lavner, J. A., and Bradbury, T. N. (2010). Patterns of change in marital satisfaction over the newlywed years. *Journal of Marriage and Family*, 72(5), 1171–1187.

Ledermann, T., and Kenny, D. A. (2017). Analyzing dyadic data with multilevel modeling versus structural equation modeling: A tale of two methods. *Journal of Family Psychology*, 31, 442–452.

Liang, M., Koslovsky, M. D., Hébert, E. T., Kendzor, D. E., Businelle, M. S., and Vannucci, M. (2021). Bayesian continuous-time hidden Markov models with covariate selection for intensive longitudinal data with measurement error. *Psychological Methods*, 28(4), 880–894.

McArdle, J. J., and Nesselroade, J. R. (2014). *Longitudinal Data Analysis Using Structural Equation Models*. American Psychological Association.

McClelland, G. H. (2000). Nasty data: Unruly, ill-mannered observations can ruin your analysis. In H. T. Reis and C. M. Judd (eds.) *Handbook of Research Methods in Social and Personality Psychology*, 1st ed. Cambridge University Press.

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward–Roger correction. *Multivariate Behavioral Research*, 52(5), 661–670.

McNeish, D., and Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods*, 25, 610–635.

McNeish, D., and Matta, T. (2018). Differentiating between mixed-effects and latent-curve approaches to growth modeling. *Behavior Research Methods*, 50(4), 1398–1414.

McNeish, D., Stapleton, L. M., and Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, 22(1), 114–140.

Matthews, R. A., Wayne, J. H., and Ford, M. T. (2014). A work–family conflict/subjective well-being process model: A test of competing theories of longitudinal effects. *Journal of Applied Psychology*, 99(6), 1173–1187.

Moller, A. C., Deci, E. L., and Elliot, A. J. (2010). Person-level relatedness and the incremental value of relating. *Personality and Social Psychology Bulletin*, 36(6), 754–767.

Nezlek, J. B. (2012). Multilevel modeling analyses of diary-style data. In M. R. Mehl and T. S. Conner (eds.) *Handbook of Research Methods for Studying Daily Life*. Guilford Press.

Orth, U., Clark, D. A., Donnellan, M. B., and Robins, R. W. (2021). Testing prospective effects in longitudinal research: Comparing seven competing cross-lagged models. *Journal of Personality and Social Psychology*, 120(4), 1013–1034.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. Sage Publications.

Reis, H. T., and Wheeler, L. (1991). Studying social interaction with the Rochester interaction record. In M. P. Zanna (ed.) *Advances in Experimental Social Psychology*, vol. 24. Academic Press.

Reynolds, B. M., Robles, T. F., and Repetti, R. L. (2016). Measurement reactivity and fatigue effects in daily diary research with families. *Developmental Psychology*, 52, 442–456.

Rights, J. D., and Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, 24, 309–338.

Schrader, S. M., Turner, T. W., Breitenstein, M. J., and Simon, S. D. (1988). Longitudinal study of semen quality of unexposed workers: I. Study overview. *Reproductive Toxicology*, 2(3), 183–190.

Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavél, F. D., Iida, M., Gleason, M. E. J., Xu, J. H., and Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences of the United States of America*, 115(1), E15–E23.

Simonsohn, U. (2018). Two lines: A valid alternative to the invalid testing of U-shaped relationships with quadratic regressions. *Advances in Methods and Practices in Psychological Science*, 1(4), 538–555.

Sin, N. L., Graham-Engeland, J. E., Ong, A. D., and Almeida, D. M. (2015). Affective reactivity to daily stressors is associated with elevated inflammation. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, 34(12), 1154–1165.

Singer, J. D., and Willett, J. B. (2003). Survival analysis. In J. A. Schinka and W. F. Velicer (eds.)

Handbook of *Psychology*, vol. 2,*Research Methods in Psychology*. John Wiley & Sons Inc.

Snijders, T. A. B., and Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd ed. Sage.

Stadler, G., Snyder, K. A., Horn, A. B., Shrout, P. E., and Bolger, N. P. (2012). Close relationships and health in daily life: A review and empirical data on intimacy and somatic symptoms. *Psychosomatic Medicine*, 74(4), 398–409.

Teague, S., Youssef, G. J., Macdonald, J. A., Sciberras, E., Shatte, A., Fuller-Tyszkiewicz, M., Greenwood, C., McIntosh, J., Olsson, C. A., Hutchinson, D., Bant, S., Barker, S., Booth, A., Capic, T., Di Manno, L., Gulenc, A., Le Bas, G., Letcher, P., Lubotzky, C. A., and the SEED Lifecourse Sciences Theme. (2018). Retention strategies in longitudinal cohort studies: A systematic review and meta-analysis. *BMC Medical Research Methodology*, 18(1), 151, https://doi.org/10.1186/s12874-018-0586-7.

Thorson, K. R., West, T. V., and Mendes, W. B. (2018). Measuring physiological influence in dyads: A guide to designing, implementing, and analyzing dyadic physiological studies. *Psychological Methods*, 23(4), 595–616.

Torre, J. B., and Lieberman, M. D. (2018). Putting feelings into words: Affect labeling as implicit emotion regulation. *Emotion Review*, 10(2), 116–124.

Uhlig, S., Meylan, A., and Rudolph, U. (2020). Reliability of short-term measurements of heart rate variability: Findings from a longitudinal study. *Biological Psychology*, 154, 107905, https://doi.org/10.1016/j.biopsycho.2020.107905.

Vajargah, K. F., and Masoomehnikbakht. (2015). Application REML model and determining cut off of ICC by multi-level model based on Markov chains simulation in health. *Indian Journal of Fundamental and Applied Life Sciences*, 5(S2), 1432–1448.

van Breukelen, G. J. P. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. *Multivariate Behavioral Research*, 48(6), 895–922.

Williamson, H. C., and Lavner, J. A. (2020). Trajectories of marital satisfaction in diverse newlywed couples. *Social Psychological and Personality Science*, 11(5), 597–604.

Wu, W., Selig, J. P, and Little, T. D. (2013). Longitudinal data analysis. In T. D. Little (ed.) *The Oxford Handbook of Quantitative Methods*, vol. 2, *Statistical Analysis*. Oxford University Press.

Zee, K. S., and Bolger, N. (2022). Physiological coregulation during social support discussions. *Emotion*, 23(3), 825–843.