CHAPTER 15

USING DIARY METHODS IN PSYCHOLOGICAL RESEARCH

Masumi Iida, Patrick E. Shrout, Jean-Philippe Laurenceau, and Niall Bolger

Diary methods involve intensive, repeated self-reports that aim to capture events, reflections, moods, pains, or interactions near the time they occur. Although modern diary methods and designs are systematic and often highly structured, they are named after the venerable tradition among literate people of making repeated, casual, and usually private notes about their own experiences, observations, attitudes, and true feelings. One of the earliest diaries of this sort is *The Pillow Book* completed in 1002 by a Japanese court lady, Sei Shonagon. In the past century, Anne Franks's diary gave the public a first-hand account of a Jewish girl's life during World War II. Historians and literary scholars have noted that diaries provide a unique perspective on eras and events, and they often use diaries as primary sources of historical record. Two of the earliest examples of diary research date back to early 1900s. *How Working Men Spend Their Time* (Bevans, 1913) tracked how individuals used their time with repeated survey design, and *Round About Pound a Week* (Pember-Reeves, 1913) examined how poor middle-class families in London used their money through repeated interview visits.

Diary methods in psychological research build on the tradition of daily written accounts and the willingness of some persons to provide exquisite detail about their experiences on a daily basis for a specified period of time. The earliest diary study in psychological research that we know of is by Csikszentmihalyi, Larson, and Prescott (1977), who examined interpersonal contacts and interaction quality among adolescents. They structured reporting forms and response intervals to make the information more systematic than free-form diaries of the literary tradition. Larson and Csikszentimihalyi (1983) called this methodology *experience sampling methods* (also called *ecological momentary assessment* [EMA]), and their method revolutionized modern psychological research by allowing investigators to capture daily experiences in participants' own, natural environment.

In the 3 decades since this first study, diary methods have been refined in many ways, notably by embracing new technology for recording events. Diary studies have become increasingly common in a variety of fields of psychology, including social (e.g., Iida, Seidman, Shrout, Fujita, & Bolger, 2008), personality (e.g., Mroczek & Almeida, 2004), clinical (e.g., Cranford, Tennen, & Zucker, 2010), developmental (e.g., Kiang, Yip, Gonzales-Backen, Witkow, & Fuligni, 2006), organizational (e.g., Butler, Grzywacz, Bass, & Linney, 2005) and health (e.g., Skaff et al., 2009) psychology. In fact, in the past 3 years, more than 250 journal articles per year have reported diary results. In addition, texts now focus explicitly on diary methods (Bolger & Laurenceau, in press; Mehl & Conner, in press).

This work was partially supported by National Institute of Mental Health Grant R01-MH60366. We thank Lesa Hoffman for sharing details of her analyses, which allowed us to generate simulation data used in the Example 3 analysis.

DOI: 10.1037/13619-016
APA Handbook of Research Methods in Psychology: Vol. 1. Foundations, Planning, Measures, and Psychometrics, H. Cooper (Editor-in-Chief)

RESEARCH QUESTIONS USING DIARY METHODS

Proponents of diary methods point to the increased ecological validity of the data, which allow a bottom-up examination of psychological processes in the participants' daily environment. Because the reports are temporally close to the experience, they also greatly reduce retrospection bias that is associated with usual survey design. In addition to these methodological advantages, diary methods allow researchers to examine questions that are not amenable in traditional study designs. These research questions can be sorted broadly sorted into three major categories: (a) What are the average experiences of an individual, and how much do the experiences vary over time? (b) Is there systematic (e.g., linear, exponential) change in experiences across days, and do such trajectories differ across persons? (c) What processes underlie a person's changes, and how do people differ in this process?

The first question often involves between-person comparisons of quantities that are summarized over time. The second and third questions have to do with descriptions and explanations of change within person. They also allow a consideration of the sequencing of different behaviors. For example, an individual might seek social support when she or he experiences a stressor, but not when life is going smoothly. To study the structural relation of support to distress, a diary researcher might compare distress levels following support receipt to levels on days when support was not available, adjusting for the severity of the stressor.

This within-persons approach is in stark contrast to cross-sectional survey designs that involve only between-persons comparisons. For example, coping researchers who use cross-sectional data might ask whether the people who use particular coping strategy also have lower levels of distress. The problem with this approach is that the within-individual associations of coping and distress are often not the same as between-individuals associations. Tennen, Affleck, Armeli, and Carney (2000) illustrated this problem with an example of the association between drinking behaviors and anxiety. Using diary methods, they show that at the between-persons level, people who drink alcohol to cope with a stressor tend to exhibit higher level of anxiety. Drinking is associated with decreased anxiety (within-person association is negative), which reinforces the behavior.

What Are the Average Experiences of an Individual, and How Much Do the Experiences Vary From Day to Day?

Many psychological phenomena are thought to operate as traits or relatively steady states. For example, attitudes, health experiences, or distress are often stable over days, if not longer. It is not uncommon for psychological measures to ask the respondent to summarize recent experience and attitudes into a single score when giving their reports. A typical example is the Dyadic Adjustment Scale (DAS) measure of relationship satisfaction symptoms (Spanier, 1976), which asks respondents to consider a statement such as, "Describe the degree of happiness . . . of your relationship," with responses that range from 0 (*extremely unhappy*) to 6 (*perfect*). Respondents are given no advice on how to weigh degree of the satisfaction or how to overcome recency biases.

With diary methods, it is easy not only to consider how happy you are from *extremely unhappy* to *perfect*, but also to consider variability in relationship satisfaction. These questions are addressed by aggregating the daily observation for each individual and calculating the means and variance over the multiple observations. This has several advantages over a retrospective assessment at one point in time. In particular, the responses obtained by diary methods, for example, daily questionnaires completed for a week, will minimize retrospective bias. Furthermore, because this approach allows for aggregation of responses, it will lead to a more valid and reliable measure.

The question on day-to-day variability often is overlooked by psychologists who are initially interested in stable traits of individuals. For example, many researchers are interested in ethnic identity as an individual difference that predicts cultural behavior, but Yip and Fuligni (2002) showed that the strength of ethnic identity varies in adolescents from

day to day, depending on where they are and who they are with. For those who become interested in within-person variability, a descriptive analysis of day-to-day variation is a necessary step to see whether further analysis is warranted. Later in this chapter we review methods that can be used to determine if observed variation can be interpreted as reliable change, or if it is simply consistent with random measurement error.

What Is the Individual's Trajectory of Experiences Across Days, and How Do Trajectories Differ From Person to Person?

These questions are what Baltes and Nesselroade (1979) have referred to as examinations of "intraindividual change and interindividual patterns . . . of intraindividual change" (p. 3). The first part of the question concerns the temporal structure of the diary data. If there is a within-person variability of the outcome of interest, how much does the passage of time explain the variability? Simple descriptions of trends over time are often called trajectories. The simplest trajectory form to consider is a linear model, which summarizes a whole series of data points with an initial level and linear change. For example, in a study of college students preparing for the Medical College Admission Test (MCAT), Bolger (1990) observed a linear increase in anxiety leading up to the examination. Although it is beyond the scope of this chapter, it is also possible to assume nonlinear change across days. For example, Boker and Laurenceau (2007) showed that relationship intimacy followed a cyclical pattern across diary period and could be fit with a sinusoidal dynamic model that required only a few parameters for each person.

The question on interindividual differences captures whether there are between-person differences in trajectories. This is often the second step of a two-step process: (a) summarize each person's trajectory with a few parameters, and (b) study the variation of those parameters across persons. For example, in the example of the MCAT study by Bolger (1990), it is possible that some participants showed sharp increases (i.e., slopes) in anxiety as they approached the exam day, whereas other participants showed very little increase in anxiety. The researchers can, then, try to explain what accounts for the differences in trajectories across days.

What Process Underlies a Person's Changes, and How Do People Differ in This Process?

The final research question is the most challenging, but it is also the most interesting question that can be asked when using diary designs. Diary designs can determine the antecedents, correlates, predictors, and consequences of daily experiences. Most of the diary studies are concerned with this question. For example, Gleason, Iida, Shrout, and Bolger (2008) examined the consequences of social support receipt on mood and relationship intimacy.

Furthermore, diary design allows for the examination of between-person differences in these processes, such that some people may show stronger within-person processes than others. Lastly, we can examine what contributes to the between-person differences. For example, Bolger and Zuckerman (1995) found that people who are high on neuroticism experienced more anxiety and depressed mood after interpersonal conflicts. The kinds of questions that can be asked of diary data continue to grow as new methodology develops. For example, Bollen and Curran (2004) described a class of statistical models called autoregressive latent trajectory (ALT) models that allow researchers to consider questions about trajectories while also considering how a process one day is especially affected by a process the day before.

TYPES OF DIARY DESIGNS

Traditional diary designs can be classified into two broad categories: time-based and event-based protocols (e.g., Bolger, Davis, & Rafaeli, 2003). In the first, data collection is scheduled or sampled according to the passage of time, and in the second, data collection is triggered by some focal experience of the participant. In addition to the traditional designs, there are innovative new designs such as device-contingent protocols. Recent developments

in technology allow for design in which participants are prompted by an electronic device on the basis of their physiological condition or surroundings.

Time-Based Design

Time-based design serves the purpose of investigating experiences as they unfold over time. These designs include *fixed-interval* schedules, where participants report on their experiences or events at predetermined intervals, and *variable-interval* schedules, where signals prompt participants to report at either random intervals or some more complicated temporal-based pattern. In both schedules, time-based designs allow researchers to examine ongoing processes that occur over a certain period.

The most common diary design is a protocol in which participants answer a series of questions about their experiences and feelings at the same time each day, but researchers increasingly consider other time-based designs that involve several reports over the course of the day. The length of the interval between assessments should be informed by the nature of the research questions. In one of the diary studies, for example, adults with anxiety disorders and their spouses were asked to report on their relationship quality at the end of the day (Zaider, Heimberg, & Iida, 2010). In another diary study, participants reported on their mood at much shorter intervals (i.e., every 3 waking hours) because investigators were interested in within-day fluctuations of positive and negative affect (Rafaeli & Revelle, 2006; Rafaeli, Rogers, & Revelle, 2007).

One of the greatest challenges of fixed time-based design is deciding the suitable spacing between the assessments. We have already that intervals depend on the research questions, but there are other important considerations. Some processes (e.g., self-perceptions of personality traits) may not change as quickly as other processes (e.g., mood); the interval can be longer for slower processes, whereas shorter intervals may be more appropriate for processes that change quickly. Another issue is that the size of the effect can vary as a function of the length of time lag between predictor and outcome of interest (Cole & Maxwell, 2003; Gollob & Reichardt, 1987). For example, in coping research, if an outcome (e.g., anxiety) is assessed a week or a month after the coping takes place, researchers may fail to capture the coping effectiveness. Even after a general spacing of observations is chosen, investigators must consider the details of implementation. For example, a research might choose to obtain assessments at a specific time of the day (e.g., 8:00 a.m., noon, 4:00 p.m.), at a specific interval (every 3 waking hours, every evening), or in a time window when it is convenient for participants.

When considering the timing of fixed interval measurements, it is important to consider the prototypic pattern of change over time and to identify which components of change are of most interest. For example, if the investigators assess cortisol every 4 waking hours, morning rise of the cortisol (Cohen et al., 2006) will not be captured. It is also possible that an important event that occurs between assessments could be missed with longer intervals. In addition, accurate recall of events or experiences becomes challenging as the interval becomes longer, and the responses might be more susceptible to biases resulting from retrospective recall and current psychological state (Shiffman, Stone, & Hufford, 2008). On the other hand, shorter intervals introduce another set of problems. Investigators may miss some effects that take longer to manifest if assessments are collected at much shorter interval (e.g., daily) than what the process or phenomenon in question unfolds (e.g., week-to-week changes). Shorter intervals also increase participant burden, therefore, investigators may need to shorten the diary period (e.g., from 4 weeks to 1 week).

The important message here is that the intervals should complement the processes that are being investigated. For example, retrospective bias may be less of a concern if investigators are interested in examining concrete, objective events (e.g., minutes exercised) rather than subjective, transient states (e.g., Redelmeier & Kahneman, 1996). Even with the increased uses of diary methodology in psychological studies, the precise timings and dynamic processes of many phenomena are largely unknown. Investigators who choose to use fixed-interval designs, however, must pick an interval before collecting any data. When the theory cannot inform how the processes or phenomenon unfold over time, Collins (2006) has suggested choosing shorter intervals.

In the variable-interval schedule, participants are asked to report their experiences at different times, and typically the participants do not anticipate when they will be asked for the next report. The investigator might use a beeper, pocket electronic organizer, cell phone, or smartphone to indicate when reports should be given, and the timing of these probes might be genuinely random or based on a pattern that appears random to the participant. In some cases, designs might be a combination of the fixed- and variable-interval designs. For example, some assessments may be collected with random beeps throughout the day, and in addition an assessment might be scheduled at the end of each day. Signaling devices for variable-interval designs have changed with emerging technology. For example, one of the earliest diary studies used an electronic paging device to transmit random beeps five to seven times per day, at which point adolescents would fill out a paper questionnaire about current activities (Csikszentmihalyi et al., 1977).

In recent years, researchers who use a variable-interval schedule tend to be interested in online assessment (e.g., how are you feeling right now?), and this type of design could minimize the retrospective recall bias. It potentially allows for random sampling of events, experiences, and behavior throughout the day. Many studies of the EMA also fall in this category, where participants report whenever they are signaled (Shiffman et al., 2008). Variable-interval schedule may be suitable for processes that are sensitive to participant expectations in which case participants automatically answer in a particular way because of the circumstances or locations (e.g., evening diary may be filled out in the bedroom right before going to bed).

One of the disadvantages of this type of design is its reliance on a signaling device, which means that the device needs to be programmed to signal at certain times. It also requires participants to carry the signaling device. Some of the issues associated with this type of design are discussed in the section Diary Format and Technology, but the main shortcoming is that it could be disruptive to the participants' daily routines, which may lead to participants avoiding to carry the device. Because participants are signaled at an interval undisclosed to them, it may also increase participant burden.

A final consideration regarding time-based diary designs is whether the time points will be considered to be distinguishable in the analysis. When all participants are surveyed at equal intervals over a fixed survey period (such as the 2 weeks before an election), then time points can be considered to be *crossed* with person. If different people are measured at different times, however, perhaps because of an event-contingent design, then the time points will be *nested* within person. Nested designs do not allow interactions of time points with other predictors to be studied as each person has a unique collection of time points. When interactions are of interest, investigators should consider using a crossed design.

Event-Based Design

When researchers are interested in rare events, such as conflicts in couples who usually are intimate and satisfied, or seizures among epileptic patients who generally are controlled, event-based designs (also known as *event-contingent* designs) are worth considering. Participants in this type of design will report every time an event meets the investigators' preestablished criterion. The most prototypical study of this kind is the Rochester Interaction Record (RIR; Reis & Wheeler, 1991). In one of their studies using RIR, Wheeler, Reis, and Nezlek (1983) asked college students to provide information on every social interaction longer than 10 minutes or more. Event-based design requires investigators to give a clear definition of the event in which they are interested. The events may go unreported or missed if the participants have ambiguous understanding of the event. One way to reduce confusion and participant burden is to choose one class of event (e.g., Laurenceau, Barrett, & Pietromonaco, 1998). Another disadvantage of this design is that there is no way to assess the compliance because it relies heavily on participants' ability to judge their situation.

Although we have presented time- and event-based designs as two separate categories, some recent studies have combined these two designs. For example, adolescents were asked to report about their environment (e.g., who they were with)

whenever they experienced self-destructive thoughts or behavior and when their handheld computer beeped randomly twice a day (Nock, Prinstein, & Sterba, 2009). Combination design can markedly improve the study design, especially if the event is extremely rare, because researchers can collect daily data even if the event does not occur or goes unnoticed by the participants.

Device-Contingent Design

In addition to traditional diary entries that require participants to stop and report on their behaviors and feelings, there are now ways to collect time-intensive data that bypass explicit participant self-report. Device-contingent design has been made possible with ubiquitous adaptations of cell phones and other electronic devices running Windows and Mac operating systems. These devices often come with a set of inputs and outputs of sensory information, such as camera devices on cell phones, microphones, Bluetooth (for wireless networking of devices over short distance), accelerometer, and global positioning system (GPS). These capabilities allow researchers to collect collateral information, such as the location where the diaries are being filled out using GPS. It is also possible to have participants wear a heart-rate monitor that is synced with a diary collection device via Bluetooth.

Potential uses and combinations are limitless, but we review one possible application of this kind of design. Suppose one is interested in coping and feelings during times of stress. With a heart-rate monitor linked to the diary device, it is possible to prompt participants to report on their experiences and situations whenever their heart rate goes beyond 100 beats per minute for a sustained period of time and to collect additional information, such as a sound recording and visual record of the participant's environment. This design has a couple of advantages. The participants do not have to detect particular instances or events (especially in event-based design) because these devices trigger when sensors detect certain situations, which, to reduce participant burden, typically can be programmed to constrain the number of times a device can trigger. More important, this allows for the continuous monitoring of physiological data with little or no awareness on the part of the participant. The designs are increasingly becoming feasible. Intille, Rondoni, Kukla, Anacona, and Bao (2003) developed a program called *context-aware experience sampling* (CAES) that allows researchers to acquire feedback from participants only in particular situations detected by sensors attached to the mobile devices.

Diary Method Design Issues

Although diary methods have important advantages over the traditional survey designs, there are some notable disadvantages as well. Some of these limitations can be minimized whereas some limitations are unavoidable. One practical concern is that most of the diary studies require a detailed training session with the participants to ensure that they understand the protocol (Reis & Gable, 2000). It is important that the participants are committed and dedicated to the participation to obtain reliable and valid data.

Diary methods often are used to capture the contexts and internal experiences of individuals as they unfold over time, so researchers ideally ask participants many questions as frequently as possible. A major obstacle of diary research is participant burden. There are three main aspects of burden: (a) length of the diary entry (e.g., 30-minute questionnaire), (b) frequency of diary responses (e.g., every 2 waking hours), and (c) length of the diary period (e.g., 9 weeks). Any one of these sources of burden can lead to subject noncompliance and attrition, and the three can have a cumulative effect on perceived burden. The challenge for any investigator is to balance the information yield with burden management. With a longer diary entry, researchers can include more questions or in-depth questions; however, less frequent responses may be more desirable to reduce participant burden. Similarly, if the participants are instructed to respond frequently (e.g., every 2 waking hours, or every time a physiological change triggers a diary entry request), researchers can closely monitor participants; however, longer diary periods may not be feasible. Broderick, Schwartz, Shiffman, Hufford, and Stone (2003) reported that using their EMA design, in which participants had to report at 10:00 a.m., 4:00 p.m., and 8:00 p.m., participant compliance

significantly dropped after the 1st week. With a longer diary period design, researchers can follow participants over a longer period and capture some slow affecting processes, but frequent reports or a longer questionnaire may lead to participant burnout. In most diary studies, lengths of the diary entry are shorter than cross-sectional surveys and this presumably allows for the collection of more frequent responses or a longer diary period. Shortening protocols requires that researchers be selective about which questions to include. Shrout and Lane (in press) warned that protocols should not be shortened by relying on only one or two items per construct because this makes it difficult to distinguish reliable change from measurement error.

Like overall sample size, the length and number of assessments in a diary design involve expenditure of resources, whether money or effort. Thus, a fundamental question is how to balance the number of subjects versus the number and length of assessments. Part of the answer to the question will come from issues of feasibility—whether subjects are available and willing to submit to the diary protocol. The other part of the answer will come from considerations of the precision of statistical estimates from the study. If the diary study is being used to obtain reliable and valid measures of between-person differences, then subject sample size will dominate the design. If the study is being used to estimate the association of events that occur within person, then the number of assessments will have to be large. We comment further about the statistical issues in the section Analysis of Diary Data.

Another important issue is the degree to which the diary reporting process changes the subject experience and behavior. Diary methods are relatively new, and there is no general theory that explains when diary completion has an impact and when it does not. Several studies show changes in participants' responses over time: For example, Iida et al. (2008, Study 1) reported that the reports of support provision increased over the 28-diary period. Several effects, such as reactance and habituation, are possible, especially if the behavior is more socially reactive. On the other hand, there is evidence that reactance does not pose a threat to validity of diary questionnaires. For example, Litt, Cooney, and Morse (1998) found that participants were more aware of the monitored behavior, but the behavior itself was not reactive. Gleason, Bolger, and Shrout (2003) also reported elevated levels of negative mood in the initial days, but the rise dissipated over 2 or 3 days. These authors argued that diaries may lead to habituation, which in turn would lead to less reactivity than other forms of survey research.

Another possibility is that participants' understanding of a particular construct becomes more complex or reliable. The repeated exposure to a diary questionnaire may enhance encoding or retrieval of relevant information. For example, in the study that observed increase in support provision (Iida et al., 2008), it is possible that participants included more behaviors that fell into the category of "social support," and it is also possible that these behaviors were more accessible at the time of diary entry. No study has directly examined this effect; however, a study by Thomas and Diener (1990) gave indirect evidence against increased complexity, at least with mood. They reported that the recall of mood did not differ following an intensive diary period. Another potential effect is that completing the diary may constrain participants' conceptualization of the domain to fit with those measured in the diary. For example, a study of social support that asks about two kinds of support (e.g., emotional and instrumental) will make participants more aware of these kinds of supportive behaviors, but they might become less sensitive to other kinds of support. Lastly, there is some evidence that a certain kind of self-reflective process has therapeutic effect (e.g., Pennebaker, 1997). This effect has not been observed with quantitative ratings, however.

A final limitation often listed for diary designs is that they produce only correlational data, and such data cannot be used to establish causal mechanisms. This limitation needs to be considered in the context of the research literature. Relative to a cross-sectional survey, diary studies allow effects to be ordered in time, and this structured data often can be used to test and reject various causal explanations. However, the same data cannot be used to establish definitively other causal explanations unless assurance can be made that the analytic model is an exact representation of the causal process, including the

time lags of causal effects. As Cole and Maxwell (2003) have shown, an incorrect specification of the timing of the causal effect (e.g., how long it takes the aspirin to reduce headache pain and keep it reduced) will lead to biased and misleading causal estimates. Thus, we conclude that relative to randomized experiments, diary studies are limited, but relative to cross-sectional studies, diary studies are a giant leap forward for causal thinking. In the future, we hope that researchers will consider supplementing diary studies with experimental designs, if that is possible, to test causes and effects more definitively.

DIARY FORMAT AND TECHNOLOGY

Once investigators decide the design of the diary study that is optimal for their research questions, they must make decisions about how to collect the data. Recent technological developments have changed the way participants report, and it is also possible to collect additional information, such as the location in which the participants make a diary entry, or to integrate these reports with physiological measures. In this section, we will review the three most commonly used diary formats: paper and pencil, brief telephone interviews, and electronic (e.g., Internet-based diary, handheld computer) diary formats.

Paper-and-Pencil Format

Early diary studies tended to use paper-and-pencil format, and this format is still one of the most widely used. It remains a good option, especially when it is coupled with a device to check the compliance (e.g., Maisel & Gable, 2009). In paper-and-pencil diary studies, participants are given packets, folders, or booklets of questionnaires, one for each diary entry, and they are instructed when to fill out and return the diary. In some studies, investigators instruct participants to return the diary every week (e.g., Bolger, Zuckerman, & Kessler, 2000), whereas in other studies, they instruct participants to return the diary within 2 to 3 days (e.g., Impett, Strachman, Finkel, & Gable, 2008, Study 2). Diary entries are similar to usual survey questionnaires, but they tend to be shorter to reduce participant burden. In general, there is little effort needed to adapt survey questions to a paper-and-pencil diary format. Another advantage of paper-and-pencil formats is that participants are familiar with the format, so it will be easier for them to fill it out. Paper-and-pencil diaries do not require complicated maintenance schedules, which is the case for some of the formats we will describe.

More and more researchers are moving away from the pencil-and-paper format because of its associated disadvantages. The main problem is participant forgetfulness with regard to compliance and diary completion. This could happen when participants fail to remember the time when they are supposed to fill out a diary entry (time-based study), or it could happen if they forget to bring the packet with them (both time- and event-based study). In either case, missed entries could lead to participants reconstructing their responses at later times, or fabricating responses. Either occurrence could undermine the advantages of the diary study. To avoid these incidences, investigators can emphasize the importance of filling out entries at specified times and that participants will not be penalized for missed entries. Asking the participants whether they filled out the entry on time may be helpful as the question suggests that some entries may not be filled out on time. It also may be useful to train the participants to use implementation intentions (Gollwitzer, 1999), which involve if–then scripts such as, "If I go to brush my teeth at bedtime, then I will complete my diary."

The bigger issue with this format is that investigators cannot track the compliance of participants without further technology. In a diary study, compliance needs to be considered both in terms of number of entries and the validity. The first one is easy to assess, but the latter is impossible to evaluate without another kind of device that tracks the compliance. For example, Rafaeli et al. (2007) asked participants to report on the time of the diary entry, and they compared their response with the information obtained from a separate computerized task. In another study, Maisel and Gable (2009) asked participants to put the entry in the envelope, seal the envelope, and stamp the date and time across the seal using an electronic stamp with a security-coded lock. Although these extra steps allow researchers to judge participant compliance, they do not ensure

that participants follow the instructions. More certainty comes from using surreptitious time-stamp devices that record the time a paper diary is opened (e.g., Stone, Shiffman, Schwartz, Broderick, & Hufford, 2002), but such devices can increase the cost of a diary study considerably.

Another limitation of paper-and-pencil format is the burden of data entry. Although this is a problem with any survey study, the problem can be pronounced because of the volume of data that are collected by the participants. For example, some of us were involved in a diary study with bar examinees and their partners (Iida et al., 2008, Study 2) that collected 44 days of diary from 303 couples. This means there were total of 26,664 ($303 \times 2 \times 44$) entries over the course of the study. During the data entry process, researchers must interpret ambiguous responses (e.g., overlapping circles on Likert-type scale responses). There may be an error during data entry process even if the responses are not ambiguous. With paper-and-pencil data records, we recommend that all of the data be entered twice by independent people, but we acknowledge that this can be costly and time consuming.

Another shortcoming is that participants might make mistakes in responses. This could happen when participants do not understand the questions or when participants miss a section of the questionnaire. This limitation is, again, not limited to diary research, but it could lead to a larger problem because of the amount of data that are collected from one individual. This can be avoided if the researchers have participants come in on the 1st day of the study.

The final shortcoming of paper-and-pencil diaries is the potential breach of confidentiality. Because the previous responses may be viewed by others in their environment, participants may hesitate to be truthful in their responses. This problem could be avoided if the participants return the diary entries more frequently (e.g., Impett et al., 2008) or are asked to seal the envelope immediately after the completion (e.g., Maisel & Gable, 2009).

Since the first structured diary study, researchers have sought to overcome the limitations of paper-and-pencil formats by using pager signaling devices, (Csikszentmihalyi et al., 1977), preprogrammed wristwatches (e.g., Litt et al., 1998), or phone calls (e.g., Morrison, Leigh, & Gillmore, 1999). Modern devices can be programmed to signal at certain times (fixed-interval schedule) or at random times (variable-interval schedule). These augmentations offer a remedy to one of the problems of paper-and-pencil format, which is participant forgetfulness. They also reduce the participant burden because they do not have to keep track of time or appropriate occasions to respond. On the other hand, these methods cannot be used for event-based sampling, and they involve additional expense.

The augmentation approach retains advantages of paper-and-pencil formats, but adds the advantage of reminding people when to fill out the questionnaires. On the other hand, participant compliance cannot be estimated by this device alone, and the cumbersome data entry remains a challenge. In addition, augmentation approaches are intrusive at times. If they are preprogrammed to signal randomly, they could go off during important meetings. This can discourage participants from carrying the signaling device, and this would defeat the purpose of the diary approach.

Brief Telephone Interviews

Another common diary data collection method is to simply call the participant using a telephone (e.g., Almeida, 2005; Waldinger & Schulz, 2010). As we discuss later in the chapter, automatic telephone systems, such as interactive voice response (IVR), also can be used to collect data. In personal telephone diary designs, trained interviewers make brief calls to the participants; the timing of the calls is determined by fixed or variable interval schedules that are set by the investigator. This data collection modality is suitable for both open-ended questions (in which participants can freely respond to a question in their own words) and questionnaires with fixed responses. The conversations can be audiotaped, or the interviewer might simply be asked to take notes and complete forms.

Brief telephone interviews have a number of advantages over paper-and-pencil dairies. One is that they can be used with persons who are not literate or who have impairments, such as visual impairments, so long as they have access to a telephone.

A second major advantage is that researchers can directly record compliance with the protocol. By actively engaging the participant, telephone interviews may help overcome participant forgetfulness and avoid any confusion about the diary protocols. In addition, researchers can allow for branching of questions (certain questions are asked depending on their previous responses), and presentation of items can be randomized to avoid habituation and boredom. If the participants provide invalid responses (out of range) or seem not to understand the questions, interviewers can correct and explain the questions. This feature of a telephone diary makes it suitable for older participants who may have trouble seeing the fonts on the diary. A key advantage of this procedure is that it involves personal interactions with the research team, and this can lead to more consistent participation and more engagement in any given interview.

Brief telephone designs also have a number of limitations. They are expensive to implement because they require hiring and training interviewers who are flexible when making calls, and who are professional in their demeanor. Almeida (2005) used a survey research center, which has professional telephone interviewers, but this can be especially costly. Participants often are not available when the interviewer makes the call, and so repeated callbacks are needed. Unless interviewers enter data directly into a computer, the same data entry costs as paper-and-pencil methods will accrue. Confidentiality may be limited especially if participants take calls at home when other family members are around, and participants may not provide honest responses to sensitive questions. Response biases may operate, and may be moderated by variables, such as ethnicity, gender of the interviewer, or other people being in the vicinity when the call is taken. The convention for most studies is to use female interviewers (ideally matched for ethnic background) because they are thought to elicit better data, and this convention is used on the basis of studies of interviewer effects in face-to-face surveys (e.g., Kane & Macaulay, 1993).

Electronic Response Formats

Electronic formats began to be used in the late 1990s (e.g., Stone et al., 1998) and have grown to be the most common design in the past decade. Uses of Internet and computer devices increased in the past decade, which also increased the comfort of participants using these formats. There are many different kinds of electronic diary data collection, but we will focus on two broad formats: fixed schedule format and variable–ambulatory assessment.

Fixed schedule formats are often implemented in diary studies that ask participants to log into a secure website and access an online questionnaire (e.g., Impett et al., 2008, Study 3). In most studies, participants are given the access code (or user name) and password to identify their data and to ensure that only one set of responses is provided. Investigators can remind the participants of the scheduled times using e-mail or phone-text messages. Some investigators provide participants with handheld devices, such as electronic personal organizers or pocket computer, which can be programmed with questions without connection to the Internet. This approach was made feasible by Barrett and Feldman-Barrett (2001), who developed a freeware diary program called ESP (Experience Sampling Program) with funding from the National Science Foundation. A third way to implement fixed schedule surveys is to ask participants to use their own phones to call a number that is associated with automatic telephone systems. These might be IVR systems that ask questions and accept verbal responses, or they might require participants to answer a few questions using their touchtone telephone pad (see Cranford et al., 2010).

Electronic response formats share many of the advantages of personal telephone interviews, and they have additional advantages. For example, they provide time stamps (and date stamps) for responses, and these give direct measures of participant compliance. By examining when responses were entered, researchers can easily identify which entries were made on time and which were not. In addition, they often allow investigators to record how long participants took to respond, and this information may be relevant to data quality and respondent burden. For example, if the participants take 3 hours to complete a diary entry that should only take 10 minutes to complete, researchers can take note.

One benefit of electronic response formats over many telephone interviews is that the responses can be easily uploaded onto the computer. This is especially true for some of the web-based questionnaires, which will put participant responses in some accessible format, such as an Excel spreadsheet or SPSS data file. This feature of electronic formats eliminates errors associated with hand entry, and it allows researchers to ensure data accuracy. Electronic formats also avoid out-of-range responses because participants are constrained to choosing a response that is available. Finally, electronic data entry avoids the contamination of the data collection by response biases and interviewer effects.

Limitations of electronic diaries are similar to those of paper-and-pencil diaries; participants must understand the diary protocol, and formatting of the web-based questionnaires must be clear. Special care must be taken to ensure that participants do not fall into a response set in which they click on responses that happen to be in the same column. Nonresponse can be a problem, but devices can be programmed to ask the respondent to check data for completeness.[1] Moreover, IVR can be scheduled to call the participant if the responses are not made within the given time frame, and researchers can contact the participants if they do not make an entry on the diary website.

Variable schedule electronic formats differ from fixed schedule formats in that participants may be asked at any time to provide information about their experience. This design requires that participants always have near them a data-entry device, such as handheld computers (i.e., palmtop computers, personal digital assistance) or a cell or smartphone. Handheld devices must be equipped with a program that allows for longitudinal data collection, such as Barrett and Feldman-Barrett's (2001) ESP software. This type of format is often coupled with variable-interval schedule designs because these devices allow for random and preprogrammed signaling. Device-contingent designs also often use this format to collect data as well. Recent technological advances, such as CAES (Intille et al., 2003), also allow for other kind of data collection, such as physiological assessment combined with typical diary questionnaires, which assesses experiences and attitudes. For researchers who are interested in assessments of fluid and transient processes, combination signaling and time-stamp responses reduce the likelihood of participant forgetfulness or retrospective recall bias.

Final Comments on Diary Format

Whichever diary format researchers choose, it is essential that careful pilot studies be carried out with participants who are drawn from the same population that will be the target of the full study. In our experience, these pilot studies almost always lead to a refinement (and improvement) of the protocol or procedure, and they help ensure that the methods are feasible with the specific population. For example, conducting a handheld computer diary study may not be ideal for older participants who have difficulty reading small text on the screen or who may lack dexterity to properly respond.

ANALYSIS OF DIARY DATA

We consider data analysis issues related to the three questions that we posed previously: (a) What are the average experiences of an individual and how much do the experiences vary from day to day? (b) What is the individual's trajectory of experiences across days, and how do trajectories differ by person? (c) What process underlies a person's changes, and how do people differ in this process? Before addressing these substantive questions, we consider the important issue of measurement quality, particularly reliability and validity. To make these statistical considerations more concrete, we focus on two substantive examples, which we describe in the following sections.

Stress and Coping During Preparation for a Professional Licensing Exam (Bar Exam Data Set)

Several of us have been involved in a large survey of support and coping in intimate couples where one partner is a recent law school graduate who is

[1]Some web-based protocols can be programmed to require a response before the participant moves on, but this option often will be in violation of informed consent assurances that say that each response is voluntary.

preparing for the bar exam. As described in various places (Iida et al., 2008, Study 2; Shrout et al., 2010), we asked both examinees and partners to complete daily paper-and-pencil questionnaires for 5 weeks before the bar exam days, 2 days during the exam, and 1 week after the examination, for a total of 44 days. We asked them to report about their general mood, relationship feelings, support transactions, troublesome events, and coping strategies each evening.

Daily Affect and Blood Glucose Levels in Diabetic Patients (Diabetes Data Set)

Skaff et al. (2009) reported results from a daily diary study of 206 diabetic patients who completed diaries for 21 days and who provided blood samples so that blood glucose could be measured. We did not have access to the original data, but we simulated artificial data that show the same pattern of results as the published paper. Not only do we use these simulated data to illustrate methods of analysis, we also describe how simulation studies such as these can be useful when planning new studies.

Psychometric Analyses of Diary Data

When reporting results from experiments or cross-sectional studies, it is considered standard good practice to report the reliability of measures and to present some evidence that they are valid. Reliability is typically defined as the tendency for measures to be able to be replicated, whereas validity is defined as evidence that a measured quantity corresponds to the theoretical construct that is the focus of the research. Shrout and Lane give details about reliability theory in Chapter 33 of this volume, and Grimm and Widaman give details about validity theory in Chapter 32 of this volume.

Standard reliability designs focus on the reliability of between-person distinctions. The reliability coefficient is interpreted to be the proportion of observed measurement variation that can be attributed to true individual difference variation. The two most common approaches to estimating reliability in psychology are test–retest designs and internal consistency designs. Test–retest designs require the investigator to arrange to have second (retest) measurements taken on persons at another occasion but before the construct of interest has changed. Under classical test theory assumptions (Crocker & Algina, 1986), a simple Pearson correlation between the test and retest provides an estimate of the reliability coefficient. Internal consistency designs allow the investigator to estimate reliability at one measurement occasion assuming that the measure is composed of two or more items. Instead of replicating the whole measurement process, the internal consistency approach asks whether different items can be considered to be replications. The most common estimate of internal consistency reliability is Cronbach's alpha. Like test–retest reliability, it measures the quality of between-person differences.

Cranford et al. (2006) showed how the internal consistency approach can be extended to the analysis of the reliability of diary data using generalizability theory (GT; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). They pointed out that diary studies typically need separate reliability analyses to describe between-person differences and within-person change. The former is an extension of the usual approach of Cronbach's alpha, but it combines information from all diary days and all scale items into a summary score for each person. The reliability of change can be separately estimated whenever there are two or more items related to a concept included in each diary. For example, it will work if participants report in separate parts of the diary how (a) angry, (b) annoyed, and (c) resentful they feel. These replicate measures allow the investigator to determine whether a participant really does have a high anger day or if some apparent daily variation might be caused by sloppy reporting.

To calculate the between-person reliability and the change reliability, the investigator uses variance decomposition software that is available in most commercial systems used by psychological researchers. Variation at the item level is broken into pieces attributable to persons, time, items, Person × Item, Time × Item, and Person × Time. Shrout and Lane (in press) provided examples of syntax for calculating these effects and for combining them into the GT reliability coefficients of Cranford et al. (2006). These methods are illustrated in the next section.

Just as two versions of measurement reliability must be considered in diary studies, so too must we

consider two versions of measurement validity. Self-report measures of fairly stable processes such as attachment style may show excellent patterns of validity in that they correlate highly with current relationship status and previous relationship difficulties, but such measures will not necessarily show validity when adapted to a daily diary research design. As Shrout and Lane (Chapter 33 of this volume) argue, a separate set of validity analyses are needed to describe how measures relate to other measures over time (rather than over people). Validity questions include face validity issues about whether participants agree that they can reflect on daily feelings related to stable self-constructs, convergent validity issues about whether daily variation in attachment feelings correlate with related constructs such as rejection sensitivity, and discriminant validity questions about whether the daily attachment measure can be shown to be distinct from (nonredundant with) related measures such as rejection sensitivity.

Multilevel Approaches to Diary Data

Many psychological measures are designed to represent *usual* levels of attitudes, behavior, affect, motivation, and so on. Some versions of measures specify a time window to consider (e.g., in the past month), whereas others make no mention of time. For example, the DAS (Spanier, 1976), a commonly used measure for relationship satisfaction, asks participants to report on various aspects of their relationships. To be specific, one of the items from the DAS asks the participants to "circle the dot which best describes the degree of happiness, all things considered, of your relationship." The response options are 0 (*extremely unhappy*) to 6 (*perfect*), with 3 (*happy*) as the midpoint. When responding, participants must mentally calculate and summarize their relationship over unspecified amount of time, and this summary may be biased by the participant's current state or situation. In a diary study, we can avoid these mental calculations by asking participants to report on their relationship satisfaction every day. We can then compute an estimate of usual satisfaction by taking the average of the daily relationship satisfaction reports by each person. These averages can be kept separate for members of the couple, or they can be further averaged to represent a couple-level satisfaction value. In addition to the mean, the within-person variability can be derived by calculating the variance of the daily relationship satisfaction reports over days by each person. Calculating these descriptive statistics can be illuminating, and can inform one about both degree and volatility of relationship satisfaction. These calculations eliminate the dependent time observations through the creation of between-person summaries of the diary experience.

More formal analysis of between and within person variation can be done using multilevel modeling (also known as hierarchical linear model, random regression modeling, and general mixed models), which is described in detail by Nezlek (see Volume 3, Chapter 11, this handbook). These methods retain the dependent data in the analysis (e.g., daily reports within person), but they explicitly model the structure of the dependent data. A number of textbooks have been written about these methods (e.g., Raudenbush & Bryk, 2002; Singer & Willett, 2003), but the approach of Raudenbush and Bryk (2002) is particularly intuitive for analysis of diary data. In their language, the analysis of the repeated observations of each person is organized around a Level 1 equation, whereas the analysis of between-person differences is organized around Level 2 equations. The beauty of these models is that they both recognize some common structure to individual life experiences and allow the investigator to consider important individual differences in manifestations of this structure.

ANALYSIS EXAMPLE 1. WHAT ARE THE AVERAGE EXPERIENCES OF AN INDIVIDUAL, AND HOW MUCH DO THE EXPERIENCES VARY FROM DAY TO DAY?

The multilevel approach is well suited to address complicated questions such as we posed in the beginning of the chapter: "What are the average experiences of an individual and how much do the experiences vary from day to day?" As we illustrate in the example that follows, the question about the average experiences can be addressed by writing a simple Level 1 (within-person) model that essentially specifies a mean across diary days and nothing

more. Level 2 (between-person) models allow the examination of individual differences such as gender and personality as well as dyadic variables such as time spent in the relationship. When we turn to questions about how the experiences vary from day to day, we build more complicated Level 1 models that describe how the participant's average experience is affected by such variables as time in the study, weekends, and even time-varying events such as conflicts or transient stressors. The participant-level experiences can then be summarized and analyzed using Level 2 models that compare time effects for males and females and so on. We illustrate the strengths of these methods in examples to follow.

Example Analysis Using Bar Exam Data Set

Our first example uses the bar exam data set described in the previous section. In this example, we explore daily relationship satisfaction reported by the examinees and partners. The relationship satisfaction was measured with two items, "content" and "satisfied," and ratings were on a 5-point scale, ranging from 0 (*not at all*) to 4 (*extremely*). The form explicitly encouraged respondents to use midpoints, and thus 11 discrete rating values are possible. We first discuss how to structure diary data. Then, we provide a description of the patterns of data, and next carry out a GT analysis of between person reliability and reliability of change in satisfaction. Finally, we apply multilevel models to describe between-person differences in the context of the diary experience.

Preliminary steps. The first step of data analysis is restructuring the data set, and this step is particularly important to diary data set. In a cross-sectional study, data set is structured such that each participant's responses are entered in a single row. Figure 15.1a

ID	content0	content1	content2	content3	content4	content5	...	content43
1	2	4	3.5	2	1.5	3	...	4
2	1	1.5	2	1.5	1	1	...	3
3	3	3	3	3.5	4	3.5	...	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
100	2	4	3.5	1	1.5	4	...	3.5

a: Wide Format

ID	Day	content
1	0	2
1	1	4
1	2	3.5
⋮		
2	1	1
2	2	1.5
⋮		
100	41	3.5
100	42	4
100	43	3.5

b: Long Format

FIGURE 15.1. Data restructuring.

shows an example such data structure (wide format), in which each row represents the means of relationship satisfaction across days (content 0–content 43). So the first person's average relationship satisfaction for the first day is 2, and this person's average satisfaction for the last day is 4. When conducting a diary analysis, it is easier to structure the data where each row represents each person's daily responses (long format); therefore, each person is going to have as many row as the diary days (see Figure 15.1b). In our current example, each participant has 44 rows of data, which means our data set consists of 4,400 rows (44 days × 100 participants).

Figure 15.2 shows a graph of four participants' trajectories of the item "content" across days. To illustrate within-person variability, we picked individuals whose average for relationship satisfaction during the diary period is 2 (*moderately*). One can see that all these participants reported higher satisfaction initially and then experienced some loss of satisfaction between Day 14 and Day 24. These are important examples of within-person change.

It is clear from Figure 15.2 that participants can differ both in level and in variability. The average score provides an efficient summary of the level and the sample variance provides a useful index of the stability of reports. When diary data are stacked (with person-time as the unit of analysis), one can use special software features such as AGGREGATE in SPSS or PROC MEANS in SAS to calculate subject-level summaries of the diary reports. The distribution of these subject mean and variance estimates can then be studied, by calculating the mean-of-means, mean variance, and the variability of these subject-level summaries in terms of standard deviations and confidence bounds. When we do this using the examinees, we get a mean of 2.72 and a standard deviation of 0.84 with 5th and 95th percentiles of 1.28 and 3.99. For partners we find a mean of 2.73 and a standard deviation of 0.78 with 5th and 95th percentiles of 1.42 and 3.88. Within-person variance can be calculated in a similar manner. In this case, we find similar amounts of within-person variability for the members of the couple; the mean variances are 0.37 for both examinees and partners.

Once these summaries are computed, we can ask questions about their mutual associations. Is there evidence that variance of satisfaction is related to the level of satisfaction in the examinee and in the partner? Do examinees with high or low average

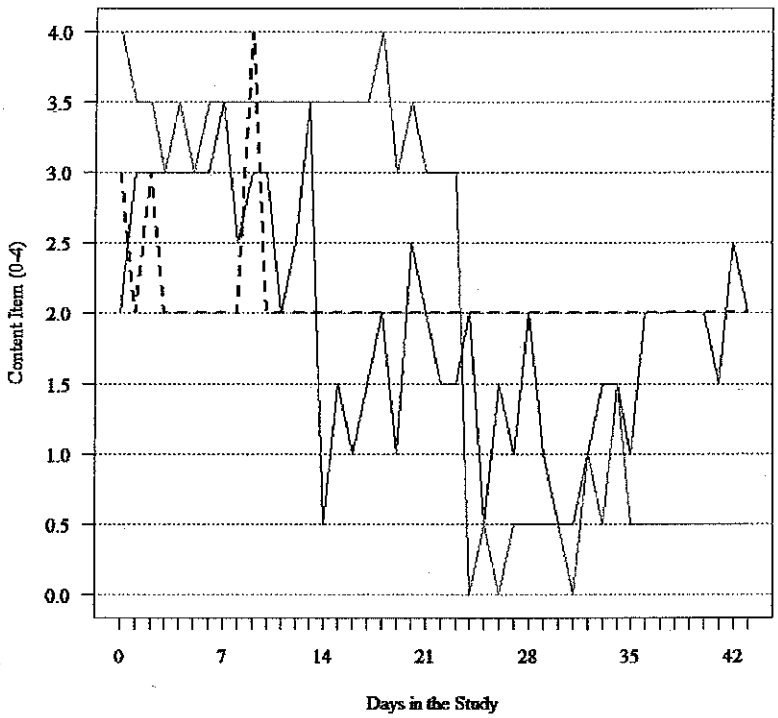

FIGURE 15.2. Trajectories of item content.

satisfaction levels have partners with similar averages? Do members of the couple tend to have similar levels of volatility of daily satisfaction during the bar exam period? These questions can be approached with simple Pearson correlations of the within person summaries. We find that there is a moderate relation between mean and variances ($r = -.29$ for examinees, and $r = -.35$ for partners). We do find substantial correlation between the mean satisfaction ratings of examinees and partners ($r = .60$), and we also find some evidence of correlation of volatility ($r = .30$).

Although these calculations are informative and easy to carry out, statisticians have suggested ways that they can be refined. The individual scores are a combination of signal and measurement error, but the above analysis treats the numbers as pure signal. Also, the simple procedure that we implemented would make use of all data, but some persons might miss a number of diary entries whereas others will make them all. This procedure does not take into account the number of days that are combined to create the summaries. Before we illustrate the multilevel statistical methods that address these issues, we first consider the question of how much error is apparent in the measurements. This requires applying the GT reliability methods of Cranford et al. (2006).

Reliability analyses using GT approach. Unlike the analyses presented so far, and the multilevel analyses that follow, the GT analysis uses variation in the item scores rather than the scale scores. Although Shrout and Lane (Chapter 33 of this volume) recommend that three or more items be used to measure important constructs, the bar exam study included only two daily items that addressed relationship satisfaction, "satisfied" and "content." On good days, both should get high scores, whereas on disappointing days both should get low scores. If the two items are discordant, the psychometric analysis concludes that error may be present.

The GT analysis states that the variability of the item scores across days and persons can be decomposed into the effects shown in Table 15.1: Person, Day, Item, Person × Day, Person × Item, Day × Item, and error. A special version of the data set was constructed in which each subject had two lines of data for each day, one with the response to "satisfied" and the other with the response to "content." These data were analyzed using the VARCOMP procedure of SAS, which uses as a default the MIVQUE method, which allows us to use the data with missing observations.[2] Table 15.1 shows the results of the variance composition analysis for partner's relationship satisfaction.

TABLE 15.1

Results of G-Study—Examinee's Relationship Satisfaction From Bar Exam Data Set

Sources of variance	Symbol	Variance estimates	Percentage
Person	σ	0.689	58.7
Day	σ	0.007	0.6
Item	σ	0.013	1.1
Person *X* Day	σ	0.281	23.9
Person *X* Item	σ	0.020	1.7
Day *X* Item	σ	0.000	0.1
Error	σ	0.166	14.1
TOTAL		1.176	100.0

In this example, three components explained most of the data. The first component is the variance due to person, the second component is the variance due to Person × Day, and third component is error variation. Variance due to person tells us that there were individual differences in the amount of relationship satisfaction the examinees reported, and it explained more than half of the variation (58.7%). Variance due to Person × Day captures individual differences on change across the study period, meaning partners had different trajectories of relationship feelings across the study, and it explained slightly less than a quarter (23.9%) of the total variance. The third component is the variance due to error, which combines the random component and the variance due to Person × Day × Item, and it explained 14% of the variance.

Cranford et al. (2006) described how the estimates of the variance components can be compared to produce a number of different reliability coefficients. These make use of different variance

[2]A similar procedure, VARCOMP, exists in SPSS. Examples of the syntax for both SAS and SPSS can be found in Chapter 33 of this volume.

components and consider the number of days (indexed by k) and the number of items (indexed by m). In this example, we focus on only two versions of reliability, the reliability of the overall mean of item responses across days (what Cranford et al., 2006, called R_{KF}),[3] and the reliability of day-to-day changes in scale scores (R_{change}). Both of these involve calculations of variance ratios, for which the numerator contains the variance of the presumed signal, either variance due to person and person by item or variance due to person by day, and the denominator contains the variance of signal plus variance due to error. According to classical test theory, the noise variation is reduced by the number of responses (m) that are averaged. This is where the impact of including additional items is most apparent.

The reliability of the average of m item scores across k days is excellent, as can be seen from the following calculation that uses Equation 4 of Cranford et al. (2006) with results in our Table 15.1:

$$R_{KF} = \frac{\sigma^2_{PERSON} + \left(\left[\sigma^2_{PERSON*ITEM}\right]/m\right)}{\left[\sigma^2_{PERSON} + \left(\left[\sigma^2_{PERSON*ITEM}\right]/m\right) + \left(\sigma^2_{ERROR}/km\right)\right]} = \frac{0.69 + (0.02/2)}{0.69 + (0.02/2) + [0.17/(44*2)]} = 0.99. \quad (1)$$

The reliability of daily change is estimated using Equation 5 of Cranford et al. (2006) along with the numerical values in our Table 15.1:

$$R_{Change} = \frac{\sigma^2_{PERSON*DAY}}{\left[\sigma^2_{PERSON*DAY} + \left(\sigma^2_{ERROR}/m\right)\right]} = \frac{0.28}{0.28 + (0.17/2)} = 0.77. \quad (2)$$

This calculation suggests that about 77% of the variance of daily change is reliable variance. Although this level of reliability is often considered to be acceptable, we can note that if we had increased the number of items to, for example, $m = 4$ items, the reliability estimate (all other things being equal) would have been 0.87. In future studies, we should take note of the possibility of improving measurement.

Multilevel analyses of diary data. Assuming the psychometric analysis suggests that there is a reliable signal in the short diary forms of measures, it is appropriate to move to substantive analyses, which are best approached using the multilevel framework introduced earlier. In a typical diary design, Level 1 units are time or event-based observations within-persons, and Level 2 units are between-persons units. Thus we say the Level 1 model is the within-person level and Level 2 is between-person level. In this chapter, we follow the notation used by Raudenbush and Bryk (2002), which distinguished the different levels rather than combining them into a reduced form equation (for alternative approach, see Fitzmaurice, Laird, & Ware, 2004).

The simplest of the multilevel equations is called a *fully unconditional* model or an intercept-only model. For the current example with examinees, the relationship satisfaction of the ith examinee at the jth time (Y_{ij}) can be represented by two equations:

$$\text{Level 1: } Y_{ij} = \beta_{0i} + \varepsilon_{ij}; \quad (3)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0i}. \quad (4)$$

In the Level 1 equation, Y_{ij}, relationship satisfaction of examinee i on day j, is represented as a function of the average over time points for each person, β_{0i}, and ε_{ij}, the deviation of relationship satisfaction on day j from the examinee's intercept. The parameter ε_{ij} can also be understood as the within-person residual, and its variance captures the within-person variance. The Level 2 equation models β_{0i}, the intercept for examinee i, as a function of grand mean, γ_{00}, and u_{0i}, the deviation of the examinee i from the grand mean. Thus, variance of u_{0i} captures the between-person variability of the average of relationship satisfaction across examinees in our data set. In multilevel modeling terms, γ_{00} is also known as a *fixed effect*, and u_{0i} is called a *random effect*. For partners, an identical model will be estimated with partners' data replacing the examinees' data.

[3] R_{KF} is named as such because it is the reliability of averages across K days for a set of fixed items. *Fixed items* mean that all participants answer the same set of items, as is the case in this example.

We can estimate this model using the MIXED procedure in SAS (syntax and partial output for examinees are available in Appendix 15.1) or other programs such as SPSS or HLM. When we do so, we get fixed effects of 2.72 for intercept for examinees and 2.73 for partners, which are identical to the estimates derived by taking the average of the averages. The variance of ε_{ij} is estimated to be 0.37 for examinees and 0.37 for partners, and the random effect of intercept is estimated to be 0.70 for examinees and 0.60 for partners. Note that the variance of the random effects from the unconditional model (0.70) is consistent with the between-person variability represented in the numerator of the formula for R_{KF} shown above. The value 0.70 is composed of the sum of overall person variance (0.689) plus one half of the person by item interaction (0.020).[4]

Assuming that the intercept estimates are normally distributed, we can estimate confidence intervals by computing a standard deviation (the square root of the variance) for each group and using the usual symmetric interval of mean ± (1.96*SD). In our study the standard deviations are 0.84 for examinees and 0.77 for partners. Using these estimates, we can calculate intervals that includes 95% of estimates ($2.72 \pm 1.96 \times 0.84$ for examinees; $2.73 \pm 1.96 \times 0.773$ for partners), which is 4.36 and 1.08 for examinees and 4.24 and 1.21 for partners. The assumption of normality is not quite correct, as is evident by the fact that the upper bounds of the confidence intervals are out of range; however, they are an approximate representation of the *spread* of intercepts in this sample.

We can build on this to look at simple between-person differences. In relationship research, we are often interested in gender differences, so we will examine how daily level of intimacy varies by the gender of the participants. To examine the gender differences, we will need to add gender as a predictor in the Level 2 question; thus the model looks as follows:

$$\text{Level 2: } \beta_{0i} = \gamma_{00} + \gamma_{01}(GENDER) + u_{0i}. \quad (5)$$

GENDER is a dummy variable that takes the value 0 for males and 1 for females. We do not find gender differences on the level of intimacy for either examinees or partners; γ_{01}(examinees) = −0.16, SE = 0.17; γ_{01}(partners) = 0.11, SE = 0.16. We can calculate 95% confidence bounds on these estimates, which are [−0.49, 0.17] for examinees and [−0.42, 0.20] for partners.

ANALYSIS EXAMPLE 2. EXAMINING CHANGES ACROSS TIME: WHAT IS THE INDIVIDUAL'S TRAJECTORY OF EXPERIENCES ACROSS DAYS, AND HOW DO TRAJECTORIES DIFFER FROM PERSON TO PERSON?

Once we know that there is sufficient within-person variance as evidenced by the variance of ε_{ij} in the previous model, the next simplest model is to examine whether passage of time explains the variance in outcome of interest, which in the current example is relationship satisfaction. Because these observations are ordered in time, this ordering may be relevant to one's analyses even if researchers have no direct interest in time. In most cases, responses from adjacent diary reports are more similar than reports farther apart. This idea is also known as the autoregressive effect. This could result if there is a change due to time. For example, marital satisfaction tends to decline after the birth of the first child (e.g., Hackel & Ruble, 1992). Autoregressive effect is also possible due to the factors not related to time. For example, fatigue could be driven by extra demands at work due to upcoming deadline.

Because we were able to examine the between-person variances of the intercept in the first research question, we can also estimate the between-person variances of trajectories. It is possible that some people show greater change compared with other people. If such between-person differences are observed, it is also possible to explain the differences. For example, people who are high in neuroticism may experience greater decline in marital satisfaction after the birth of the first child.

[4]Some programs, such as HLM, report an intraclass correlation from the multilevel model, and this is identical to R_{KF} for the fully unconditional model.

Example Analysis Using Bar Exam Data Set

We will continue to use the daily relationship satisfaction scale for this example analysis, but we will focus on partners' report for this example. Figure 15.3 shows the average of relationship satisfaction across participants for 44 days of the study, where day 35 is the first day of the examination. On average, the marital satisfaction reported by partner steadily decreases as the examinees approach the examination and starts to increase right before the examination. Formal conclusion must be withheld until we estimate a statistical model, however, because the figure is a representation of mean anxiety across participants for each day of the diary period, which ignores the within-person differences.

Modeling a linear trajectory. To formally test the within-person change in satisfaction, we must represent the multilevel model of individual change over time. We will first start with the simplest model of trajectory. Equations 6 through 8 are often referred to as *individual growth models* (Singer & Willett, 2003, pp. 49–51):

$$\text{Level 1: } Y_{ij} = \beta_{0i} + \beta_{1i}(DAY0)_j + \varepsilon_{ij}; \quad (6)$$

$$\text{Level 2: } \beta_{0i} = \gamma_{00} + u_{0i}; \quad (7)$$

$$\text{Level 2: } \beta_{1i} = \gamma_{10} + u_{1i}. \quad (8)$$

In the Level 1 equation, Y_{ij}, represents satisfaction of partner i on Day j, and *DAY0* is an indicator of the sequence of days, with the first day coded as 0. In the equation, satisfaction is modeled as a function of intercept, β_{0i}, which is the average satisfaction for partner i on Day 0, β_{1i}, which represents the linear change of anxiety for each successive day, and ε_{ij}, the deviation of satisfaction on day j from the partner's intercept after accounting for the linear effect of *DAY0*. The effects of day, in the current data set, is also a proxy variable for increased stressor because elapsed time also means approaching examination. The first Level 2 equation represents β_{0i}, the intercept for partner i on Day 0, as a function of γ_{00}, the mean of satisfaction on Day 0 after adjusting for linear change (a fixed effect), and u_{0i}, the deviation of the partner i from the mean (a random effect). The second Level 2 equation models β_{1i}, time slope for partner i, as a function of γ_{10}, fixed effect of the time slope, and u_{1i}, the deviation of the partner i from the average time slope (random effect of *DAY0*). In this model, because there are two random

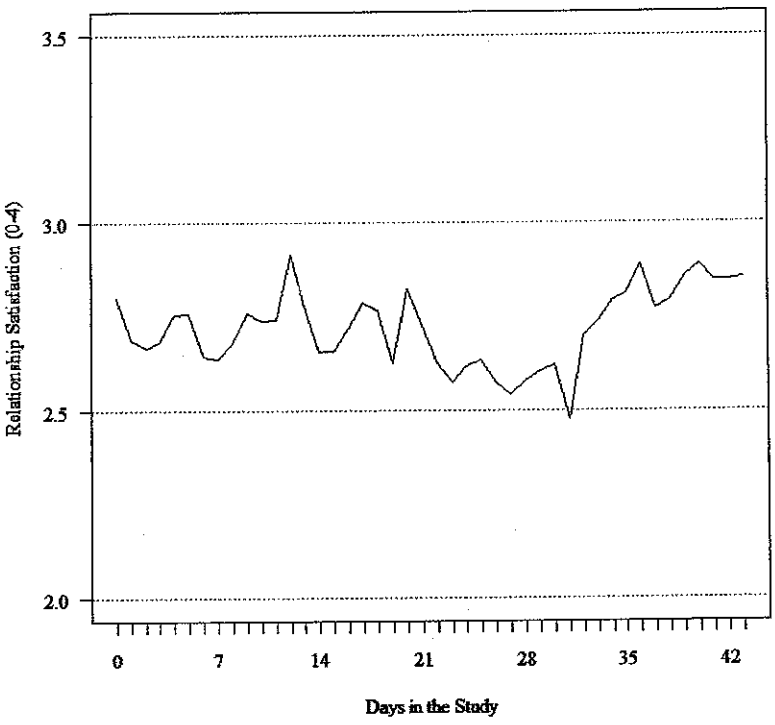

FIGURE 15.3. Trajectory of average relationship satisfaction for partners.

effects, we can ask whether they are correlated. When random effects are positively correlated, it implies that as initial level increases, the slope also increases (e.g., partners who have high initial level of satisfaction also increases more with each passage of day); negative correlation means that as the initial level increases, the slope decreases (e.g., people who have high initial level of satisfaction show a decline in satisfaction across days).

Another important detail when modeling time is the specification of residual error structures ($\varepsilon_{ij,}$), also known as the residual variance covariance matrix, because there is a statistical consequence for not specifying a proper structure (Greene & Hensher, 2007). In diary data, *residual error* refers to the unexplained variance associated with the particular day, and we expect errors to be correlated over time *within* people. There are several ways to structure the residual variance covariance matrix; however, we focus only on autoregressive structure. In an autoregressive model, the errors are structured as follows:

$$\begin{bmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \sigma^2\rho^3 & \cdots \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \cdots \\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 & \sigma^2\rho & \cdots \\ \sigma^2\rho^3 & \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \quad (9)$$

The autoregressive model is relatively efficient because it estimates only two parameters, σ and ρ. For those who are interested in other types of residual structures, please see Singer (1998).

In the example at hand, the fixed effects of the intercept is 2.76, which implies that the average relationship satisfaction on Day 0 (see Appendix 15.2). The fixed effect of Day is –0.001, which means that satisfaction decreases over the entire study period by approximately 0.001 unit each day, but in this case, the effect is not significant. The random effect of intercept is estimated to be 0.52, so if we take the square root of the estimate ($\sqrt{0.52} = 0.72$), it will give us the standard deviation of intercepts in our sample. Again, we can calculate the 95% confidence interval of intercepts, and we get 1.35 and 4.17. Similarly, the random effect of day is estimated to be 0.0001, which corresponds to a standard deviation of 0.01, and the 95% confidence interval of linear change is –0.02 and 0.02.

Thus far, we have not paid attention to the significance test of either fixed effects or random effects. The significance tests of fixed effects are tested by *t* test with degree of freedom approximated by Satterthwaite estimates, which recognize that several different variances are being estimated from the same data (Raudenbush & Bryk, 2002). The significance tests of random effects are slightly more complex. The Wald *z* test used to test random effects are known to be conservative, and methodologists recommend that differences in the deviance (–2 log likelihood, or –2LL) be used instead (Singer & Willett, 2003, pp. 116–117). In our example, the –2LL of full model is 7876.3, and the –2LL of the model without the random intercept is 8211.6. Therefore, the likelihood ratio difference is 335.3, and this value is significant at $p = .001$ level using the chi-square test with degree of freedom of 1. Similarly, the random effect of *DAY* is significant at $p = .001$ level, but remember that the fixed effect of *DAY* is not significant. In other words, there is no linear change in the relationship satisfaction *on average*, but there is significant variation around this effect, such that some people show a slight decline (–0.02 calculated in the preceding paragraph) yet other people show a slight increase (0.02) across the study period. The covariance between the random effects of intercept and effect of *DAY* are not significant. Because the random effects of intercept and *DAY* are significant, we know that there is systematic variation in these effects, which will be explored later.

Modeling a quadratic trajectory. The pattern in Figure 15.3 suggests a curvilinear trajectory of relationship satisfaction across days, so we will now test the quadratic effect of within-person change in satisfaction. The following equations capture the quadratic trajectory of satisfaction:

$$\text{Level 1: } Y_{ij} = \beta_{0i} + \beta_{1i}(DAYc)_j + \beta_{2i}(DAYc^2)_j + \varepsilon_{ij}, \quad (10)$$

$$\text{Level 2: } \beta_{0i} = \gamma_{00} + u_{0i}, \quad (11)$$

$$\text{Level 2: } \beta_{1i} = \gamma_{10} + u_{1i}, \text{ and} \quad (12)$$

$$\text{Level 2: } \beta_{2i} = \gamma_{20} + u_{2i}. \quad (13)$$

This model is similar to the prior set of equations, but with an inclusion of the squared day effect in Level 1 equation and additional equation in Level 2. Another difference from the prior equation is that *DAYc* is now centered on 15th day, which means that the 15th day is coded as 0, whereas *DAY0* was centered on the first day of the study in the linear trajectory model. The 15th day corresponds to 2 weeks before the examination. β_{2i} is the estimate for quadratic effect, and it is modeled as a function of γ_{20}, fixed effect of the quadratic effect and u_{1i}, the deviation of the examinee *i* from the average quadratic effect (random effect of $DAYc^2$).

Table 15.2 summarizes the results of this quadratic trajectory analysis. The interpretation of intercept changes from the previous model. Instead of representing the adjusted value on the first day, and it is now the adjusted average level of satisfaction on Day 15, which is 2.69. The estimate of day also changes from previous model, and this is the linear change of satisfaction on Day 15. Therefore, marital satisfaction is decreasing by 0.009 unit on Day 15.

TABLE 15.2

Results of Analysis Examining Quadratic Change of Relationship Satisfaction

Effects		
Fixed effects	γ[a]	***SE***
Intercept (level on Day 10)	2.688**	0.081
DAYc (linear change)	−0.009**	0.003
$DAYc^2$ (quadratic change)	0.001**	0.0001
Random effects[b]	τ	**LR**
Level 2 (between-person)		
Intercept	0.627***	762.9
DAY (linear change)	0.001***	45.1
$DAYc^2$ (quadratic change)	0.000	1.0
Level 1 (within-person)		
Autocorrelation	0.297***	438.7
Residual	0.333[a]	NA

Note. LR = likelihood ratio; NA = not applicable.
[a]The model without Level 1 residual variance is implausible; therefore, the deviance difference cannot be calculated. [b]Because the model that had covariances of random effects was unstable, we only estimated the random effects.
**$p < .01$. ***$p < .001$.

We also find that $DAYc^2$ is significant, which suggests that the satisfaction follows a quadratic pattern.

Moderation effect. We can explore how individuals vary in the linear and quadratic effects. This is especially useful when we have a theory about how to explain systematic individual differences. In our example, we chose relationship closeness as measured by Inclusion of Others (IOS) scale (Aron, Aron, & Smollan, 1992) as a potential source of the individual difference (e.g., moderator). IOS is a single-item, pictorial measure of relationship closeness that ranges from 1 (*two circles are barely touching*) to 7 (*two circles are highly overlapped*). We thought that persons who included their partner as part of their self-concept would be more influenced by the partner's stressful bar exam experience. To examine the moderating effect of IOS, we ran a cross-level interaction model, where centered IOS (IOSc) was included as predictors in Level 2 equations. The Level 1 equation remained the same as prior model, and the Level 2 equations were as follows:

$$\text{Level 2: } \beta_{0i} = \gamma_{00} + \gamma_{01}(IOSc) + u_{0i}, \quad (14)$$

$$\text{Level 2: } \beta_{1i} = \gamma_{10} + \gamma_{11}(IOSc) + u_{1i}, \text{ and} \quad (15)$$

$$\text{Level 2: } \beta_{2i} = \gamma_{20} + \gamma_{21}(IOSc) + u_{2i}. \quad (16)$$

The first Level 2 equation models β_{0i}, the intercept for partner *i* on Day 0, as a function of γ_{00}, fixed effect of intercept; γ_{01}, fixed main effect of IOSc; and u_{0i}, the deviation of the partner *i* from the grand mean. In all of these models, IOSc was centered around 5.19, the mean in the sample. The second Level 2 equation models β_{1i}, the time slope for partner *i*, as a function of γ_{10}, fixed effect of the time slope; γ_{11}, fixed moderating effects of IOS of time on relationship satisfaction; and u_{1i}, the deviation of the partner *i* from the average time slope (random effect of DAYc). The third Level 2 equation models β_{2i}, the quadratic effect for partner *i*, as a function of γ_{20}, fixed effect of the quadratic effect; γ_{11}, fixed moderating effects of IOS of quadratic effect on relationship satisfaction; and u_{1i}, the deviation of the partner *i* from the average time quadratic effect (random effect of $DAYc^2$).

We find that IOS is a significant between-person predictor of marital satisfaction such that people who are high on IOS tend to have higher levels of

anxiety on the 15th day of the study by 0.33 units ($\gamma_{01} = 0.33$, $SE = 0.05$). Therefore, this gives some evidence that IOS explains the variability (random effects) across participants in our sample. IOS was also a marginally significant moderator of the quadratic effect ($\gamma_{03} = -0.0002$, $SE = 0.0001$), such that partners who were higher on IOS tended to show a smaller quadratic effect (see Figure 15.4), which suggests that their daily marital satisfaction is less affected by the impact of the bar examination. As for the day effect, IOS does not moderate the effect ($\gamma_{02} = 0.002$, $SE = 0.002$). In other words, IOS does not explain the variation in the effects of day across partners.

ANALYSIS EXAMPLE 3. PREDICTING CHANGE AND DAILY PROCESS: WHAT PROCESS UNDERLIES A PERSON'S CHANGES, AND HOW DO PEOPLE DIFFER IN THIS PROCESS?

Examining changes across time is interesting, but psychological researchers may be more interested in the predictors of the changes. Diary methods are a great way to model the proposed causal relationship as a temporal within-person process if the temporal measurement corresponds to when causes and effect change. Diary design does not permit inferences as strong as that of experimental designs, however, because experimental conditions are experienced in temporal order, which leaves open the possibility of (a) carryover effect from the previous experiences; (b) order effects, in which the particular order of experiences moderates the effects of interest; or (c) expectancy effects, in which previous experience changes the meaning of subsequent experience.

There are four ways in which diary methods can strengthen causal inferences. One is that diary studies differentiate between-person and within-person associations. Many within-person variables collected in diary data, such as mood, vary both within and between persons. Within-person associations often are referred to as using the person as his or her own control, and it allows us to treat results as pertaining to the relationship between within-person changes in X and Y. In practical terms, we can differentiate

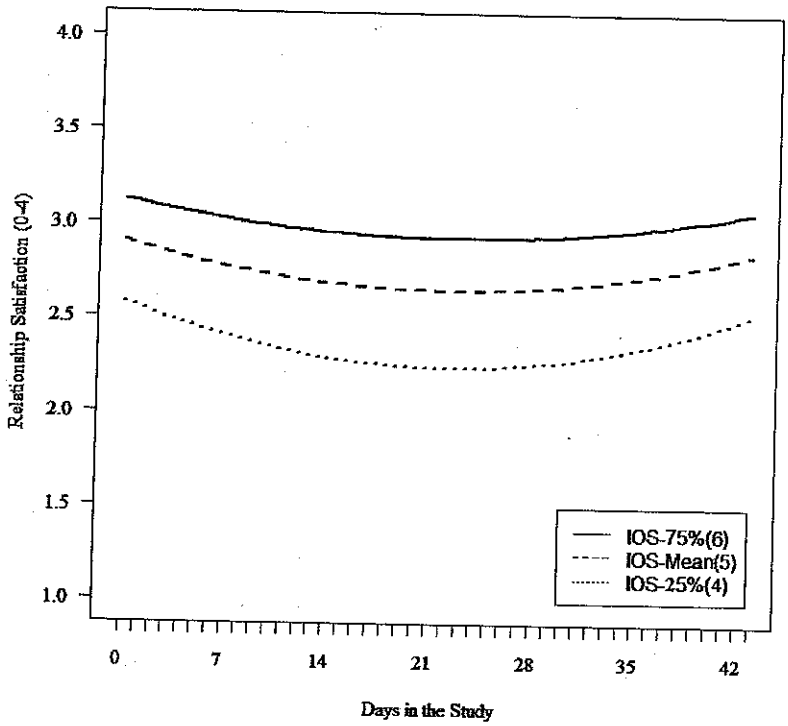

FIGURE 15.4. Quadratic effect of relationship satisfaction by IOS showing that higher IOS scores are associated with larger quadratic effects than lower IOS scores.

two sources of variation by including each person's mean value of *X*. Second is by including elapsed time because there could be an effect of time even if that is not the main research questions. For example, if the diary period is longer, participant boredom or habituation may affect the levels and interrelationships among variables. Third, analysis should accurately reflect the temporal structure of cause–effect relationship, which also includes having measurements at accurate times. Last, additional confounding within-person variables need to be taken into account. These could include lagged variables of *Y* that could be acting as an alternative explanation of the relationships that one is examining.

Example Analysis Using Health Psychology Data

To illustrate how we can examine the process underlying how a person changes, we use the simulated health psychology data on the basis of the study by Skaff et al. (2009). In this data set, we have negative affect (*NA*) and waking blood glucose from 207 Type 2 diabetic patients for 21 days. As in the previous two examples, we use multilevel modeling to examine the association between prior day affects and waking blood glucose measure. Before we run any analysis, it is important to calculate means of negative and positive affect for each person, and they will be used to derive the within-person centered variable of negative affect, which in turn allows us to estimate the within-person effect. As described, this approach allows us to differentiate the within- and between-person associations. The within-person effects can be understood as the effects of the negative affect that is greater than the person's average.

The model is represented in the following equations:

$$\text{Level 1: } Y_{ij} = \beta_{0i} + \beta_{1i}(NA_{j-1} - MeanNA_i) + \varepsilon_{ij}, \quad (17)$$

$$\text{Level 2: } \beta_{0i} = \gamma_{00} + \gamma_{01}(MeanNA)_i + u_{0i}, \text{ and} \quad (18)$$

$$\text{Level 2: } \beta_{1i} = \gamma_{10} + u_{1i}. \quad (19)$$

In the Level 1 equation, Y_{ij}, blood glucose of person i on day j, is modeled as a function of intercept, β_{0i} (the average blood glucose for person i), β_{1i} (within-person effect of *NA* from previous day), and ε_{ij} (the deviation of negative mood on Day j). The lagged effects of *NA*, β_{1i}, addresses the main research questions in this example—the within-person associations of previous day negative affect and waking blood glucose. In other words, when the participants experience negative affect more than they usually do, how much the blood glucose changes the following morning.

The Level 2 equation models β_{0i}, the intercept for person i on Day 0, as a function of γ_{00}, fixed effects of intercept (grand mean of blood glucose on Day 0); γ_{01}, between-person effect of mean negative affect (*MeanNA*) for person i; and u_{0i}, the deviation of the person i from the grand mean. The second Level 2 equation models β_{1i}, effect of *NA* for person i, as a function of γ_{10}, average effect of *NA* for all participants in this sample (fixed effect of *NA*), and u_{1i}, the deviation of the person i from the average effect of *NA* (random effect of *NA*).

Table 15.3 summarizes the results of the analyses examining affect and blood glucose. The fixed effect of the intercept is 4.900, which is the estimate of the average blood glucose on the first day of the study (Day 0). The fixed within-person effect of NA is 0.022, which means that when participants

TABLE 15.3

Results of Analysis Predicting Blood Glucose

Effects		
Fixed effects	γ^a	***SE***
Intercept (level on Day 0)	4.900**	0.018
Negative affect (previous day)	0.022**	0.004
Mean negative affect (between-person)	0.085**	0.030
Random effects	τ	**LR**
Level 2 (between-person)		
Intercept	0.064***	26.8
Negative affect (previous day)	0.001	0.0
Intercept—negative affect covariance	−0.001***	20.2
Level 1 (within-person)		
Residual	0.030[a]	NA

Note. LR = likelihood ratio; NA = not applicable.
[a]The model without Level 1 residual variance is implausible; therefore, the deviance difference cannot be calculated.
**$p < .01$. ***$p < .001$.

experience more negative affect than usual, their glucose increases by approximately 0.02 unit the following day. The between-person effect of mean negative affect is 0.085, which means that individuals who, on average, have high negative affect across the diary period tend to have higher waking blood glucose.

DIARY EXTENSIONS

Diary methods provide us with rich data on psychological processes as they unfold. The statistical analyses we have reviewed so far are only a starting point. More advanced methods are available in Bolger and Laurenceau (in press) and Mehl and Conner (in press) as well in the current literature. In the next section, we mention some recent developments and speculations about future directions.

Dynamic Systems Model

One of the ways diary data can be examined is by using dynamic systems models, also known as dynamical systems models (see Volume 3, Chapter 16, this handbook). Dynamic systems, generally defined, are self-contained sets of elements that interact in complex, often nonlinear ways to form coherent patterns, with an underlying assumption that these systems regulate themselves over time. One of the key concepts is the idea of stationary attractor points (or equilibrium state), a set of points that regulates the system. The idea of stationary attractor points translates well into psychological terms; it can be thought of as the norm or the average state. When the oscillation can be described by a sine or cosine function, a mathematical property states that the rate of acceleration of the function (second derivative) is a linear combination of the rate of change (first derivative) and level of the function (see Boker, 2001, for a detailed explanation). Gottman, Murray, Swanson, Tyson, and Swanson (2002) applied more complicated dynamic models to investigate marital interaction in their book *The Mathematics of Marriage: Dynamic Linear Models*. Dynamic models can be appealing analytic techniques especially for researchers who collect physiological data because these data tend to exhibit cyclical patterns and abundant data are collected from the individual. Applications dynamical systems modeling (second-order linear oscillator modeling) to dyadic diary data can be found in Boker and Laurenceau (2006, 2007).

Categorical Variables

In this chapter, we have focused on analysis for continuous outcomes; however, many of the variables that psychologists are interested in may be categorical or counts (e.g., whether support was received, whether a conflict occurred, number of alcoholic beverages consumed). For these types of outcomes, the analytic strategy described in this chapter will lead to misspecified models. Fortunately, however, a number of appropriate alternative methods can be considered, including nonlinear multilevel models. These are readily available in statistical packages like HLM and SAS (GLIMMIX and NLMIXED procedures). For details on these types of multilevel analyses, please see Bolger and Laurenceau (in press) and Hox (2010).

Multivariate Multilevel Analysis

Diary researchers typically ask participants to report on a variety of behaviors, feelings, and attitudes over time and are interested in how these processes operate in a multivariate system. For example, Gleason et al. (2008) wondered how the costs of daily support on anxiety could be reconciled with the benefits of daily support on relationship closeness. In the context of multilevel analyses, they used a multivariate approach that has been described in detail by Raudenbush and Bryk (2002). This approach involves a treating the different outcomes as if they were repeated measures, and the data on the different outcomes are stacked together. A special provision is needed, however, to recognize that each outcome has its own set of fixed and random effects. Multivariate systems can also be considered in the context of structural equation methods. Bollen and Curran (2004) described a class of models called ALT models that will be of special interest to researchers who have relatively few diary measurements.

Diaries in Dyads, Families, and Groups

One of the most growing types of diary is the diary in which more than one person from a dyad (e.g.,

married couples), families (e.g., parents and eldest children), and groups (e.g., students in classrooms) participate in the study. In these types of data, there are two different sources of nonindependence: Observations are repeated within persons, and persons are nested within dyads, families, or groups. To account for both types of nonindependence, it is sometimes useful to use three-level multilevel models for groups greater than three people. For the smaller groups (dyads and three-member family groups with prespecified roles), a two-level multilevel model is sufficient in which the lowest level represents the multivariate repeated measures mentioned in the previous section (Bolger & Laurenceau, in press; Bolger & Shrout, 2007; Laurenceau & Bolger, 2005). Other books have been dedicated to analyses of nonindependent data (e.g., Kenny, Kashy, & Cook, 2006).

Simulation as a Tool for Diary Researchers

When planning a diary study, how can one determine the sample size of persons and time points, and what is the trade-off between increasing either kind of *n*? Many diary researchers first carry out a statistical simulation study, using software such as SAS or Mplus (Muthén & Muthén, 2010). These simulations create data that are like what you hypothesize the process to be, and analyses of these artificial data give the researcher some idea of how small the standard errors get as the number of participants or time points increase. To carry out a simulation, one needs to go through six steps: (a) Write down a Level 1 statistical model for the within-subject responses, (b) write down a series of Level 2 statistical models that describe how the subjects differ in their within-subject process, (c) consult the literature to identify plausible effect sizes for the fixed effects, (d) consult the literature to identify plausible values to represent variability of within- and between-subject processes, (e) write a program to simulate data according to the model from Steps a through d, and (f) analyze the simulated data to determine how precise the results are. We have provided an example of simulated health data. Although we had access to an interesting published study (Skaff et al., 2009), we did not have access to the original data. From the published results and some personal communication with one of the authors, we were able to generate data that resembled daily glucose and mood patterns. We provide this simulation syntax as an example online (see https://sites.google.com/a/asu.edu/using-diary-methods-in-psychological-research).

If one hopes to carry out diary studies with participants who have some rare condition, it may be difficult to recruit more than a dozen or so subjects. Not only will statistical power be challenging in this case, but also the usual inferential methods of multilevel models may be misleading. Much of the statistical theory for multilevel models assumes large samples. In cases such as this, simulation studies can be used to study the impact of sample size on both power and the usual control of Type I error. If small sample data are available, one could use simulation methods to carry out resampling studies of those data. For an example of a simulation-based power analysis for diary data, please see Bolger, Laurenceau, and Stadler (in press).

CONCLUSION

Although relatively new to psychology, diary study designs are changing the way psychologists think about psychological process. They help survey workers determine when retrospective memory is problematic, and when it can be counted on. When spaced over months, they can provide invaluable information about development in youth and adolescents. Diary accounts provide new sources of compelling data for health psychologists, and they allow stable individual differences to be distinguished from meaningful change. We predict that the number of diary studies will continue to accelerate in the literature.

Nonetheless, there are many ways that diary designs will continue to be refined. Technological advances, particularly in microrecording devices, will open new doors for noninvasive measurement. Statistical methods will continue to be developed to deal with the complicated temporal patterns in nonstationary data. New methods for small samples and for complicated dependent data will be proposed. For those interested in making an impact on research methodology, the area of diary studies is fruitful ground.

Appendix 15.1
Syntax for SAS

```
PROC MIXED covtest noclprint;
CLASS couple;
MODEL content = /s;
RANDOM int / SUBJECT = couple TYPE = vc;
RUN;
```

PARTIAL OUTPUT

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
Intercept	couple	0.6986	0.1005	6.95	< .0001
Residual		0.3706	0.007	993	46.37	< .0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t value	PR > |t|
Intercept	2.7189	0.08409	99	32.33	< .0001

Appendix 15.2
Syntax for SAS

```
PROC MIXED covtest noclprint DATA = dchap4;
CLASS couple day;
MODEL pcontent = day0 /s;
RANDOM int day0 / SUBJECT = couple TYPE = UN
REPEATED day/SUBJECT = couple TYPE = ar(1);
TITLE "Linear Trajectory Analysis - Partners";
RUN;
```

PARTIAL OUTPUT

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	couple	0.5177	0.08232	6.29	< .0001
UN(2,1)	couple	0.000153	0.001163	0.13	0.8950
UN(2,2)	couple	0.000126	0.000032	3.94	< .0001
AR(1)	couple	0.3442	0.01592	21.63	< .0001
Residual		0.3581	0.009352	38.29	< .0001

Fit Statistics
-2 Res Log Likelihood 7876.3
AIC (smaller is better) 7886.3
AICC (smaller is better) 7886.3
BIC (smaller is better) 7899.3
Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	2.7588	0.07606	99	36.27	< .0001
day0	-0.00116	0.001492	99	-0.78	0.4402

References

Almeida, D. M. (2005). Resilience and vulnerability to daily stressors assessed via diary methods. *Current Directions in Psychological Science, 14*(2), 64–68. doi:10.1111/j.0963-7214.2005.00336.x

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology, 63*, 596–612. doi:10.1037/0022-3514.63.4.596

Baltes, P. B., & Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 1–39). New York, NY: Academic Press.

Barrett, L. F., & Feldman-Barrett, D. J. (2001). An introduction to computerized experience sampling in psychology. *Social Science Computer Review, 19*, 175–185. doi:10.1177/089443930101900204

Bevans, G. E. (1913). *How workingmen spend their time* (Unpublished doctoral thesis). Columbia University, New York, NY.

Boker, S. M. (2001). Differential models and "differential structural equation modeling of intraindividual variability." In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 5–27). New York, NY: Oxford University Press. doi:10.1037/10409-001

Boker, S. M., & Laurenceau, J.-P. (2006). Dynamical systems modeling: An application to the regulation of intimacy and disclosure in marriage. In T. A. Walls & J. L. Schafer (Eds.), *Models for intensive longitudinal data* (pp. 195–218). New York, NY: Oxford University Press.

Boker, S. M., & Laurenceau, J.-P. (2007). Coupled dynamics and mutually adaptive context. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 299–324). Mahwah, NJ: Erlbaum.

Bolger, N. (1990). Coping as a personality process: A prospective study. *Journal of Personality and Social Psychology, 59*, 525–537. doi:10.1037/0022-3514.59.3.525

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology, 54*, 579–616. doi:10.1146/annurev.psych.54.101601.145030

Bolger, N., & Laurenceau, J.-P. (in press). *Diary methods*. New York, NY: Guilford Press.

Bolger, N., Laurenceau, J.-P., & Stadler, G. (in press). Power analysis for intensive longitudinal measurement designs. In M. R. Mehl & T. Conner (Eds.), *Handbook of research methods for studying daily life*. New York, NY: Guilford Press.

Bolger, N., & Shrout, P. E. (2007). Accounting for statistical dependency in longitudinal data on dyads. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 285–298). Mahwah, NJ: Erlbaum.

Bolger, N., & Zuckerman, A. (1995). A framework for studying personality in the stress process. *Journal of Personality and Social Psychology, 69*, 890–902. doi:10.1037/0022-3514.69.5.890

Bolger, N., Zuckerman, A., & Kessler, R. C. (2000). Invisible support and adjustment to stress. *Journal of Personality and Social Psychology, 79*, 953–961. doi:10.1037/0022-3514.79.6.953

Bollen, K. A., & Curran, P. J. (2004). Autoregressive Latent Trajectory (ALT) models: A synthesis of two traditions. *Sociological Methods and Research, 32*, 336–383. doi:10.1177/0049124103260222

Broderick, J. E., Schwartz, J. E., Shiffman, S., Hufford, M. R., & Stone, A. A. (2003). Signaling does not adequately improve diary compliance. *Annals of Behavioral Medicine, 26*, 139–148. doi:10.1207/S15324796ABM2602_06

Butler, A. B., Grzywacz, J. G., Bass, B. L., & Linney, K. D. (2005). Extending the demands-control model: A daily diary study of job characteristics, work-family conflict and work-family facilitation. *Journal of Occupational and Organizational Psychology, 78*, 155–169. doi:10.1348/096317905X40097

Cohen, S., Schwartz, J. E., Epel, E., Kirschbaum, C., Sidney, S., & Seeman, T. (2006). Socioeconomic status, race, and diurnal cortisol decline in the coronary artery risk development in young adults (CARDIA) Study. *Psychosomatic Medicine, 68*, 41–50. doi:10.1097/01.psy.0000195967.51768.ea

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112*, 558–577. doi:10.1037/0021-843X.112.4.558

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology, 57*, 505–528. doi:10.1146/annurev.psych.57.102904.190146

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin, 32*, 917–929. doi:10.1177/0146167206287721

Cranford, J. A., Tennen, H., & Zucker, R. A. (2010). Feasibility of using interactive voice response to monitor daily drinking, moods, and relationship processes on a daily basis in alcoholic couples. *Alcoholism: Clinical and Experimental*

Research, 34, 499–508. doi:10.1111/j.1530-0277-.2009.01115.x

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, and Winston.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Csikszentmihalyi, M., Larson, R., & Prescott, S. (1977). Ecology of Adolescent Activity and Experience. *Journal of Youth and Adolescence*, 6, 281–294. doi:10.1007/BF02138940

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley-Interscience.

Gleason, M. E. J., Bolger, N. P., & Shrout, P. E. (2003, February). *The effects of study design on reports of mood: Understanding differences between cross-sectional, panel, and diary designs*. Poster presented at the annual meeting of the Society for Personality and Social Psychology, Los Angeles, CA.

Gleason, M. E. J., Iida, M., Shrout, P. E., & Bolger, N. P. (2008). Receiving support as a mixed blessing: Evidence for dual effects of support on psychological outcomes. *Journal of Personality and Social Psychology*, 94, 824–838. doi:10.1037/0022-3514-.94.5.824

Gollob, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development*, 58, 80–92. doi:10.2307/1130293

Gollwitzer, P. M. (1999). Implementation intentions—Strong effects of simple plans. *American Psychologist*, 54, 493–503. doi:10.1037/0003-066X.54.7.493

Gottman, J. M., Murray, J. D., Swanson, C. C., Tyson, R., & Swanson, K. R. (2002). *The mathematics of marriage: Dynamic nonlinear models*. Cambridge, MA: MIT Press.

Greene, W. H., & Hensher, D. A. (2007). Heteroscedastic control for random coefficients and error components in mixed logit. *Transportation Research Part E: Logistics and Transportation Review*, 43, 610–623.

Hackel, L. S., & Ruble, D. N. (1992). Changes in the marital relationship after the first baby is born: Predicting the impact of expectancy disconfirmation. *Journal of Personality and Social Psychology*, 62, 944–957. doi:10.1037/0022-3514.62.6.944

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Mahwah, NJ: Erlbaum.

Iida, M., Seidman, G., Shrout, P. E., Fujita, K., & Bolger, N. (2008). Modeling support provision in intimate relationships. *Journal of Personality and Social Psychology*, 94, 460–478. doi:10.1037/0022-3514.94.3.460

Impett, E. A., Strachman, A., Finkel, E. J., & Gable, S. L. (2008). Maintaining sexual desire in intimate relationships: The importance of approach goals. *Journal of Personality and Social Psychology*, 94, 808–823. doi:10.1037/0022-3514.94.5.808

Intille, S. S., Rondoni, J., Kukla, C., Anacona, I., & Bao, L. (2003, April). *A context-aware experience sampling tool*. Paper presented at the CHI '03 Extended Abstracts on Human Factors in Computing Systems, Fort Lauderdale, FL.

Kane, E. W., & Macaulay, L. J. (1993). Interviewer gender and gender attitudes. *Public Opinion Quarterly*, 57(1), 1–28. doi:10.1086/269352

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York, NY: Guilford Press.

Kiang, L., Yip, T., Gonzales-Backen, M., Witkow, M., & Fuligni, A. J. (2006). Ethnic identity and the daily psychological well-being of adolescents from Mexican and Chinese backgrounds. *Child Development*, 77, 1338–1350. doi:10.1111/j.1467-8624.2006.00938.x

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. *New Directions for Methodology of Social and Behavioral Science*, 15, 41–56.

Laurenceau, J.-P., Barrett, L. F., & Pietromonaco, P. R. (1998). Intimacy as an interpersonal process: The importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges. *Journal of Personality and Social Psychology*, 74, 1238–1251. doi:10.1037/0022-3514.74.5.1238

Laurenceau, J.-P., & Bolger, N. (2005). Using diary methods to study marital and family processes. *Journal of Family Psychology*, 19, 86–97. doi:10.1037/0893-3200.19.1.86

Litt, M. D., Cooney, N. L., & Morse, P. (1998). Ecological Momentary Assessment (EMA) with treated alcoholics: Methodological problems and potential solutions. *Health Psychology*, 17, 48–52. doi:10.1037/0278-6133.17.1.48

Maisel, N. C., & Gable, S. L. (2009). The paradox of received social support: The importance of responsiveness. *Psychological Science*, 20, 928–932. doi:10.1111/j.1467-9280.2009.02388.x

Mehl, M. R., & Conner, T. S. (Eds.). (in press). *Handbook of research methods for studying daily life*. New York, NY: Guilford Press.

Morrison, D. M., Leigh, B. C., & Gillmore, M. R. (1999). Daily data collection: A comparison of three methods. *Journal of Sex Research*, 36, 76–81. doi:10.1080/00224499909551970

Mroczek, D. K., & Almeida, D. M. (2004). The effect of daily stress, personality, and age on daily negative affect. *Journal of Personality*, 72, 355–378. doi:10.1111/j.0022-3506.2004.00265.x

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Authors.

Nock, M. K., Prinstein, M. J., & Sterba, S. K. (2009). Revealing the form and function of self-injurious thoughts and behaviors: A real-time ecological assessment study among adolescents and young adults. *Journal of Abnormal Psychology, 118*, 816–827. doi:10.1037/a0016948

Pember-Reeves, M. (1913). *Round about a pound a week.* London, England: Bell.

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science, 8*, 162–166. doi:10.1111/j.1467-9280.1997.tb00403.x

Rafaeli, E., & Revelle, W. (2006). A premature consensus: Are happiness and sadness truly opposite affects? *Motivation and Emotion, 30*, 1–12. doi:10.1007/s11031-006-9004-2

Rafaeli, E., Rogers, G. M., & Revelle, W. (2007). Affective synchrony: Individual differences in mixed emotions. *Personality and Social Psychology Bulletin, 33*, 915–932. doi:10.1177/0146167207301009

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

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66*, 3–8. doi:10.1016/0304-3959(96)02994-6

Reis, H. T., & Gable, S. L. (2000). Event-sampling and other methods for studying everyday experience. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 190–222). New York, NY: Cambridge University Press.

Reis, H. T., & Wheeler, L. (1991). Studying social-interaction with the Rochester Interaction Record. *Advances in Experimental Social Psychology, 24*, 269–318. doi:10.1016/S0065-2601(08)60332-9

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4*, 1–32. doi:10.1146/annurev.clinpsy.3.022806.091415

Shrout, P. E., Bolger, N., Iida, M., Burke, C. T., Gleason, M. E. J., & Lane, S. P. (2010). The effects of daily support transactions during acute stress: Results from a diary study of bar exam preparation. In K. T. Sullivan & J. Davila (Eds.), *Support processes in intimate relationships* (pp. 175–200). New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780195380170.003.0007

Shrout, P. E., & Lane, S. P. (in press). Psychometrics. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life.* New York, NY: Guilford Press.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23*, 323–355. doi:10.3102/10769986023004323

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford University Press.

Skaff, M. M., Mullan, J. T., Almeida, D. M., Hoffman, L., Masharani, U., Mohr, D., & Fisher, L. (2009). Daily negative mood affects fasting glucose in Type 2 diabetes. *Health Psychology, 28*, 265–272. doi:10.1037/a0014429

Spanier, G. B. (1976). Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads. *Journal of Marriage and the Family, 38*, 15–28. doi:10.2307/350547

Stone, A. A., Schwartz, J. E., Neale, J. M., Shiffman, S., Marco, C. A., Hickcox, M., . . . Cruise, L. J. (1998). A comparison of coping assessed by ecological momentary assessment and retrospective recall. *Journal of Personality and Social Psychology, 74*, 1670–1680. doi:10.1037/0022-3514.74.6.1670

Stone, A. A., Shiffman, S., Schwartz, J. E., Broderick, J. E., & Hufford, M. R. (2002). Patient noncompliance with paper diaries. *British Medical Journal, 324*, 1193–1194. doi:10.1136/bmj.324.7347.1193

Tennen, H., Affleck, G., Armeli, S., & Carney, M. A. (2000). A daily process approach to coping: Linking theory, research, and practice. *American Psychologist, 55*, 626–636. doi:10.1037/0003-066X.55.6.626

Thomas, D. L., & Diener, E. (1990). Memory accuracy in the recall of emotions. *Journal of Personality and Social Psychology, 59*, 291–297. doi:10.1037/0022-3514.59.2.291

Waldinger, R. J., & Schulz, M. S. (2010). What's love got to do with it? Social functioning, perceived health, and daily happiness in married octogenarians. *Psychology and Aging, 25*, 422–431. doi:10.1037/a0019087

Wheeler, L., Reis, H., & Nezlek, J. (1983). Loneliness, social-interaction, and sex-roles. *Journal of Personality and Social Psychology, 45*, 943–953. doi:10.1037/0022-3514.45.4.943

Yip, T., & Fuligni, A. J. (2002). Daily variation in ethnic identity, ethnic behaviors, and psychological well-being among American adolescents of Chinese descent. *Child Development, 73*, 1557–1572. doi:10.1111/1467-8624.00490

Zaider, T. I., Heimberg, R. G., & Iida, M. (2010). Anxiety disorders and intimate relationships: A study of daily processes in couples. *Journal of Abnormal Psychology, 119*, 163–173. doi:10.1037/a0018473