Longitudinal Designs

Aidan G.C. Wright
University of Pittsburgh

and

Kristian E. Markon
University of Iowa

This chapter was peer-reviewed and accepted for publication in the American Psychological Association’s (APA) Handbook of Clinical Psychology, J.C. Norcross, G.R. VandenBos, & D.K. Freedheim (Eds.). The APA holds the copyright for this chapter. This article may not exactly replicate the final version published in the handbook. It is not the copy of record.

Please address correspondence to Aidan G.C. Wright, Department of Psychology, University of Pittsburgh, 4121 Sennott Square, 210 S. Bouquet St., Pittsburgh, PA 15260. Email: aidan@pitt.edu
Longitudinal Designs

Longitudinal designs have become a critical tool in clinical psychology to address numerous questions at a variety of time scales and levels of explanation. As such, longitudinal designs can be approached in a number of different ways. In this chapter, we define longitudinal studies, present their core principles and applications, and review several landmark studies and achievements. We conclude with potential limitations of longitudinal studies and probable future directions. The overarching aim of this chapter is to provide a conceptual framework for those thinking of longitudinal designs, and more generally, for those considering the use of time to better understand psychological processes.

This chapter focused on conceptual issues and findings in longitudinal studies, but we caution readers interested in pursuing a longitudinal design to simultaneously consider the theoretical constructs of interest, their operationalization in sampling schedule and instrumentation, and their quantitative articulations in formal statistical models. The most powerful and successful longitudinal designs will tightly interweave these strands into the fabric of their study.

Description and Definition

A longitudinal study can be defined as one where multiple measurements are taken across two or more time points. Through repeated measurements longitudinal studies enlist time to estimate stability, change, or other features of unfolding processes. Time is leveraged to make inferences about processes of interest, such as in using the ordering of events across time to better understand causality; using an interval of time to understand the stability of a psychological construct within the dynamics of development; or using repeated measurements across time to understand how different patterns of growth predict outcome.
Longitudinal studies encompass a wide variety of designs, depending on the number of measures at each time point, the number of time points, and the interval between time points. For example, micro-level intensive data collection protocols over a short time period, such as hourly or daily intervals over the course of days, characterize experience sampling, day reconstruction, and daily diary studies. On the other end of the spectrum are macro-level or long-term protocols, such as developmental and aging studies that employ intervals of several years. In between are studies typically occurring over the course of weeks or months, such as test-retest or follow-up studies.

Longitudinal designs may also differ in their univariate versus multivariate focus. We use these terms here loosely to refer to univariate studies whose focus is on single constructs or measures repeatedly observed across time, or multivariate studies whose focus is on relationships among multiple constructs or measures repeatedly observed across time. Examples of univariate studies include those examining the stability of personality traits across age (e.g., Roberts & DelVecchio, 2000), changes in cognitive ability over development (e.g., Schaie, 1994), variability in mood over the course of days or weeks (e.g., Penner, Shiffman, Paty, & Fritzsche, 1994), or variation in growth trajectories of externalizing psychopathology (e.g., Miner & Clarke-Stewart, 2008). Examples of multivariate studies include cross-lagged designs aiming to disentangle the direction of causality in associations between parenting constructs and psychopathology (Burt, McGue, Krueger, & Iacono, 2005), profile stability (Donnellan et al., 2007), as well as studies whose goal is understanding idiographic constructs and measurement processes explaining relationships among measures for single individuals as they are observed across time (e.g., person-specific factor analysis of within-person correlations; Borkenau & Ostendorf, 1998; Nesselroade, Gerstorf, Hardy, Ram, 2007).
Longitudinal studies can be either variable-centered or person-centered (Molenaar & Campbell, 2009). That is, longitudinal studies can either have as a principle goal identifying patterns in constructs, or in characterizing patterns within individuals. For example, a variable-centered approach to development of personality pathology might focus on changes in narcissism with age; a person-centered approach might focus on individual differences in the development or expression of narcissism.

**Principles and Applications**

**Time as a Variable**

The defining feature of longitudinal study is that it enlists time as a variable via repeated measurements. *Yet we are rarely interested in time per se but rather what it represents.* Indeed, the power of time as a variable is that it can stand in any equation or model we have for psychological processes without having to know those processes exactly. The usual impetus for conducting a longitudinal study is to uncover or better understand a process. But frequently little may be known about its properties. Initially, even basic features such as the process’ time scale, effect size, or whether it serves to maintain stability, contribute to structured change, or unstable fluctuations, might be unknown. Early exploratory work might serve to clarify these parameters.

Time itself only offers evidence for a process; it is not the process per se. For instance, the finding that, on average, personality disorder criteria decline at the rate of 1.4 criteria per year in early adulthood (Lenzenweger, Johnson, & Willett, 2004) implies at least one process that leads to this outcome. Time stands in as that unknown process (or processes). Subsequently, additional variables (e.g., psychological treatment, personality traits, relationship status) can be entered as predictors of the decline to begin to uncover the nature of that process.

This logic can be extended to the sampling frame or partitioning of data sets, such that
participants can be identified that are known to differ in their fundamental processes, even when it is unclear what these may be. To illustrate, it has been argued that antisocial behavior in adulthood can be distinguished by its first appearance in childhood or adolescence (Moffitt et al., 1996). Differences in the temporal manifestation of conduct problems suggest a different mechanism across groups. The underlying mechanism need not be fully explicated for these groupings to be fruitfully culled from longitudinal data, and used as predictors, outcomes, or in other ways based on their timing of onset.

**Scale of Measurement Intervals**

Although time can stand in for unknown processes, studies should be designed around a time-scale on which the key processes of interest are thought to play out. Important in any study is carefully selecting a sampling schedule based on the phenomenon of interest (Collins, 2006). If one wants to powerfully leverage time, knowing the right timing of the processes of interest will make for a more powerful fulcrum in the study. Certain phenomena are known to emerge slowly over the course of years (e.g., cortical development in adolescence and early adulthood), others over months (e.g., human gestational period), others over a week (e.g., fluctuations of daily mood), and still others shift from moment to moment (e.g., affective responses to situational stressors).

Macro-level change is usually what comes to mind when discussing longitudinal studies, but recently, there has been increased emphasis on using intensive repeated measurement designs in order to capture micro-level psychological processes (Moskowitz et al., 2009). Referred to variously as ambulatory assessment, experience sampling methodology, and ecological momentary assessment, the goal is to repeatedly measure individuals with short delays between measurement (for example, hours to a day). Although the use of daily diaries or daily
questionnaires is not novel, recent years have seen a dramatic increase in clinical psychology of studies designed around sampling on the order of multiple times per day, separated by only an hour or a few between samplings (Trull & Ebner-Priemer, 2013). This has allowed for powerful tests of basic assumptions about certain forms of psychopathology (e.g., affective instability in borderline personality pathology; Trull et al., 2008) and more nuanced articulations of complex relationship processes (e.g., Sadikaj et al., 2013).

**Number of Measurements**

Closely related to the scale of measurements is their total number. Decisions about number of measurements are mostly related to statistical issues that go beyond the chapter, but we note a couple of issues here for consideration. When trajectories of change are of interest, at least three waves of data is necessary, but four or more is advisable in order to increase reliability, and required if non-linear change is hypothesized (e.g., quadratic growth curves). Furthermore, when sampling certain behaviors that are low base-rate (e.g., self-harm), sufficient numbers of assessments need to be conducted in order to capture the behavior with enough frequency to fit statistical models.

**Temporal Constructs of Interest**

Most temporal features of interest can be thought of in terms of two dimensions. The first of these is which aspect of a construct’s distribution across time is of interest—for example, whether mean level (progression) or variance (instability) across time is the focus. The second is the level of aggregation across individuals at which changes across time are being described or explained—for example, whether changes are being examined at the level of individual persons or in aggregate, across persons.

Many questions focus on mean-level trends over time, either overall in a group or at the
level of individuals. Decreases in cognitive ability, for example, are evident in mean-level changes in the population; however, these changes in level also differ across individuals, such that some individuals decline more quickly than others, and others still might actually increase somewhat across periods of adulthood. In both cases, the focus is on changes in level across time, either at the group or individual level.

Other questions focus on variability in constructs over time. At a basic level, for example, one might be interested in diurnal fluctuations in mood, and how that changes seasonally in the population over the course of the year. Other questions pertain to individual differences in variability—for example, whether individuals who are more variable in their mood during the week exhibit greater psychopathology than those who are less variable in their mood (e.g., Eid & Diener, 1999; Murray, Allen, & Trinder, 2002). Many traditional questions about stability in rank ordering of individuals (e.g., correlations between personality at one age and another) can also be recast as questions about individual variability in change across time. For example, the correlation between individuals' levels of psychopathology at two time points can be construed in terms of the variance of within-person changes in psychopathology across time (Rodgers & Nicewander, 1988).

**Measurement Across Time and Persons**

As in any research, measures in a longitudinal study must be valid. Unlike cross-sectional studies, longitudinal studies raise the possibility that measurement processes may change with time like the underlying constructs they reflect. This possibility poses potential challenges of interpretation, in that any observed change over time may reflect either changes in the constructs or changes in the measurement process. For this reason, it is important to evaluate such changes in measurement across time.
Longitudinal studies of cognition provide good examples of modeling possible changes in measurement. Measures of cognitive ability demonstrate well-established trends with age, generally increasing from childhood to adulthood, then declining throughout adulthood, especially in domains of speeded and nonverbal processing (e.g., McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002). These declines in adulthood can be parsed into age as well as cohort effects (i.e., the Flynn effect, wherein later-born cohorts produce larger scores on cognitive measures; Flynn, 1999; see Chapter 4).

Although substantive explanations have been provided for these age and cohort effects, it has also been argued that changes in measurement over time can also partially explain these phenomena. For example, measures of cognition are susceptible to practice effects, which mask actual decline over time (Ferrer, Salthouse, Stewart, & Schwartz, 2004; Salthouse, 2010). These practice effects are larger in younger individuals, leading to exaggerated observed declines with age (Salthouse, 2010). Other changes in measurement properties of cognitive tests across cohorts also partially explain cohort-related changes in cognitive measures over time (Schaeie, et al., 1998; Wicherts, et al., 2004). Such findings imply that at least part of the Flynn effect can be attributed to changes across time in how the measures function.

Similar age-related changes in response styles and factor structures have been demonstrated for personality measures as well. There is evidence, for example, that acquiescent responding (“yeah saying”) declines from childhood to adulthood (Soto, John, Gosling, & Potter, 2008). In another example, lack of sexual interest is a strong marker of schizoid pathology in young adults, but provides little information about the trait in older adults for whom libido is normatively diminished (Balsis et al., 2007). Overall, these phenomena highlight the importance of being mindful that measurement processes as well as constructs may change with time.
Longitudinal studies also highlight possible differences in measurement processes across persons. Response styles and biases differ from individual to individual, constructs may be more or less salient for one individual relative to another, and individuals otherwise differ in how they perceive and respond to measures for reasons that have little or nothing to do with the underlying construct. Modern psychometric theory (e.g., item response theory) often assumes that longitudinal data on intraindividual processes are not available and that interindividual data can be used as a substitute or approximation (Molenaar, 2008). This assumption is not warranted in general, however, as the structure of interindividual and intraindividual variation is not necessarily the same, and longitudinal information about response processes can be obtained.

Individual variation is most clearly revealed through repeated measurements of the same individuals over time. One central question in much of psychometric theory is “what response would you expect a given individual to have to a given measure?” (Holland, 1990). In longitudinal designs, one can examine how specific individuals approach a given measure differently. For example, although personality traits may have a familiar Big Five structure when considered across persons, that structure is not necessarily evident in each individual's responses within person across time (Borkenau & Ostendorf, 1998).

**Scale of Change and Observation Intervals**

Interpreting the nature of change or stability in constructs of course depends on the time scale being examined. For constructs not expected to change over the interval being examined (e.g., personality or cognitive ability over a day interval, for normal settings), change might be interpreted as error, and stability as reliability or dependability. For constructs that might change significantly over the time scale being examined (e.g., mood over a day interval, or personality over the course of several years), change might be interpreted as meaningful. In this way,
identifying an appropriate interval at which to collect observations is critical: if the interval between observations is too short, changes are unlikely to be meaningful; if they are too long, it becomes difficult to model components of change due to transient, random influences.

Accounting for transient or time-specific effects becomes critical in deciding what measures are collected together at any given time point, because the effect of time of observation can sometimes dwarf other relationships that might be of interest. For example, investigators examining structural or measurement questions, such as questions surrounding the factor or latent structure of a set of measures, need to be careful to either collect data on such measures at the same timepoints, or at intervals that are close enough in time that the effect of time of observation is not significant. Otherwise, the risk of spurious “wave factors” emerges, in which factors reflect time of observation rather than important factors influencing response such as mood, personality, or attitudes. This issue has plagued some measures in the large scale National Epidemiological Study of Alcoholism and Related Conditions, for example, where spreading different personality measures across different time points conflated construct domain and time of assessment (Trull, Vergés, Wood, Jahng, & Sher, 2012).

Choosing an appropriate interval is also likely to depend on the principal questions of the study in addition to the constructs being examined. For example, making inferences about causality requires designs with intervals appropriate to the putative causal and effect variables involved. In particular, it is necessary to choose intervals such that change in a causal variable can be modeled along with the hypothesized change in the effect variables. In a longitudinal study examining the effect of parent behavior on externalizing psychopathology in children, for example, an interval that is too short might miss changes in either parent behavior or externalizing; if it is too long, changes in externalizing due to parenting changes might get
washed out by other developmental trends. Similarly, a study hoping to examine the effects of acute versus chronic stressors would require time scales appropriate to both effects—one might consider a burst design, for example, in which sets of observations at short intervals are set apart in longer intervals (see e.g., Sliwinski, 2008). Ultimately, a good theory of the scale of change for the target construct is needed to design an effective study (Collins, 2006).

**Sampling of Persons**

Another important consideration is who to include in a longitudinal study. As always, the ability to generalize from the study findings to a population of interest depends on the sample. In longitudinal designs a critical question is whether sample characteristics will give an unrepresentative view of patterns of change in the population. In clinical psychology, often this results from associations between initial values on constructs of interest and subsequent change (e.g., the law of initial value, or ceiling and floor effects). For example, individuals who are extremely high on a measure of psychopathology initially may have little room to increase, and may be much more likely to decline in their levels of psychopathology over time (cf. Vaidya, Gray, Haig, Mroczek, & Watson, 2008). Samples that are restricted to such individuals may provide misleading evidence regarding changes in the population as a whole, possibly overestimating declines. However, when the purpose of a study is to understand rates of stability and change of pathology among those diagnosed with the disorder, as might be the case to inform questions of prognosis, then recruiting clinically elevated samples is likely advantageous. Recognizing the populations to which the study results generalize or do not is imperative.

Relatedly, when the variables used for selecting participants in to the study and the tracking change over time are identical, it creates issues with interpretation of the results. Specifically, the independent and dependent variable become one and the same, raising issues of
“endogeneity” when considering the trajectories of change over time (Duncan et al., 2004). To protect from this, the use of selection procedures that differ from the repeated measures procedures should ideally be employed. Furthermore, initial participant characteristics that are not the primary focus of the study may be associated with other features of change. For example, elevations in negative emotion tend to be associated with greater variability in mood over time (e.g., Eid & Diener, 1999; Penner, et al., 1994); such patterns suggest that restricting samples to individuals very high in negative emotion may lead to inflated estimates of variability in mood over time, or that conversely, restricting samples to individuals very low in negative emotion may suggest unrealistically low variability in mood over time.

Characteristics of the persons being sampled have practical implications for how studies are conducted as well. Depending on the sample and constructs being studied, for example, it may be more or less difficult to collect certain forms of data, or collect data at certain times. For instance, populations that are transient, are difficult to contact, have periods of inaccessibility (e.g., cannot respond to prompts during work hours), or are otherwise highly unreliable may pose challenges for longitudinal designs. However, these types of considerations can be mitigated through rigorous collection of contact information, intensive follow-ups, and appropriate incentives, and past studies have achieved excellent compliance even when studying putatively highly unreliable samples, such as homeless individuals with cocaine use problems (Freedman et al., 2006). Thus, challenging populations do not ipso facto preclude intensive protocols, although investigators need to consider these issues when designing studies.

**Limitations of the Approach**

As with any type of design, researchers working with longitudinal studies face a number of potential difficulties, which can be ameliorated through prevention and awareness.
Longitudinal designs afford more nuanced and complex inferences about psychological phenomena than cross-sectional observational studies, but these affordances come at a certain cost.

**Costs of Longitudinal Designs**

With multiple measurement occasions, all other things being equal, longitudinal designs obviously cost more in resources than cross-sectional designs simply because more observations are being collected on the same number of people. Whether financial or otherwise in nature—the ultimate financial cost per participant, the time and effort involved in tracking and retaining participants, and so forth—the resources required are generally greater than a cross-sectional design that is comparable to one wave of the longitudinal design.

**Tradeoffs in Observation and Measurement**

In addition, with multiple waves of data collection, the burden on participants can also be greater, which in turn can constrain the type of data collected. The repetition of measures effectively creates a tradeoff between collecting more measures fewer times, or fewer measures more times. For example, in an intensively longitudinal design, such as one might encounter in an EMA study, it would be infeasible to collect data on hundreds of items several times daily. The frequency of data collections would necessitate many fewer measurements per observation, either by reducing the number of different instruments, or by reducing the number of indicators per instrument. In this way, increasing the density of observations per person in a longitudinal study can constrain either the diversity of constructs being examined or the reliability of each measure. Especially when considering intensive longitudinal designs, there is the potential for participant fatigue (i.e., response rate diminishes over time), or reactivity (i.e., the participants alter their behavior due to having to report on it so frequently).
On the other hand, with repeated measurements of constructs for which little change is expected over the time frame of the study, it is possible to increase the dependability of the measurements (i.e., reliability in a cross-time sense). In this way, longitudinal studies can also engender compromises between different forms of reliability, specifically internal consistency reliability versus test-retest reliability. Emerging research on these two forms of reliability underscores that they are distinct, in that measures relatively reliable in one sense are not necessarily reliable in the other sense (Chmielewski & Watson, 2009; Watson, 2004).

Dependability appears to vary by construct domain (e.g., measures of extraversion are more dependable than measures of agreeableness; Chmielewski & Watson, 2009), and measures that are more affectively laden tend to be less dependable (Anusic, Lucas, & Donnellan, 2012; Wood & Wortman, 2012). Future research on different forms of reliability and the factors influencing them will undoubtedly help inform decisions involving the selection of measures in longitudinal designs.

**Poor Choice of Measurement Interval**

Above we suggested that it is important for investigators to be thoughtful about selecting the measurement interval (i.e., the distance in time between measurements). It is worth emphasizing that a poor choice of interval may not only diminish the potential to find real change effects, but may lead to spurious findings or opposite results (Pelz & Lew, 1970). Depending on the functional form of the change process, especially when change is non-linear, certain lag intervals may provide results of opposite sign. One way to guard against this possibility is to pilot measurement lags of different temporal distance or sample more frequently in order to test for effects across lags of different size.

A related problem is when investigators are allured by the illusion of meaningful effects
simply because measurements are separated by time. This issue most often arises in the estimation of certain types of longitudinal mediation models. Consider the case where moderately correlated constructs (e.g., Neuroticism, Depression, Anxiety) are sampled across three time points over which these constructs are reasonably stable. Within each wave the measures will exhibit virtually identical correlations, and across waves the measures exhibit a similar but slightly attenuated pattern of associations. The investigator then may test a mediation model wherein the effect of Neuroticism at Time 1 on Depression at Time 3 is mediated by Anxiety at Time 2. Although a reasonable model, it is likely to be confirmed simply because these constructs have stable relationships and one measure (i.e., Anxiety) is sample more proximally to the outcome. However, a highly similar pattern is likely to emerge by shuffling the ordering of the constructs in the mediation model, and largely reflects their contemporaneous associations. Thus, including time between measurements does guarantee interesting results.

**Measurement Noninvariance**

Finally, as noted earlier, it is important to account for time-related measurement changes (measurement noninvariance) in designs, to avoid concluding that change has occurred in constructs of interest when the change in observed variables is due to changes in response process over time. If older individuals approach a measure of impulsivity differently from younger individuals, for example, observed changes might not reflect actual changes in impulsivity with age, but differences in what is relevant to assessing impulsivity in younger versus older individuals. Such concerns were prominent in discussions of the DSM-5 Attention Deficit Hyperactivity Disorder (ADHD) criteria, where there has been some uncertainty about how much adult declines in ADHD-related behavior are due to changes in ADHD per se, or in the way it is assessed.
Similar concerns might arise at the level of individual differences as well, in that if different individuals or groups of individuals (e.g., different sexes or sociocultural populations) are responding to instruments differently, one might conclude that the groups differ in their change trajectories when in fact they only differ in their response processes. If a measure functions better in one group than in another, for example, it might also be more sensitive to change over time. In such a situation, it would be misleading to conclude that observed changes between groups reflect actual changes in the constructs of interest: are smaller changes in one group due to smaller true changes or differences in measurement sensitivity across groups?

On the surface it is difficult to envision any limitation with collecting data over multiple waves as opposed to a single measurement. However, there are very real practical challenges one faces when sampling over time (e.g., cost, infrastructure needed for participant tracking), as well as several insidious potential threats to the validity of the study and conclusions drawn if one is not mindful of the challenges involved in longitudinal designs.

**Landmark Studies**

That which constitutes a *landmark* longitudinal study can be difficult to define, and the studies one believes should make this list likely will vary by an individual’s area of interest. Undoubtedly, our own interests and proclivities influenced those studies we selected. Nevertheless, we offer up a list of studies that are easily defensible as landmarks due to their ambition, scope (both in size and length of follow-up), and their contribution to new knowledge.

- *Genetic Studies of Genius* (Terman & Oden, 1959): Lewis Terman’s study of 1528 children (age 11) of superior intelligence (M IQ ~ 150), followed for many decades by Terman himself and subsequently by additional investigators into old age.

- *Deviant Children Grow Up* (Robins, 1966): A naturalistic follow-up study that tracked down a
large sample (~500) of individuals 30 years after initial referral to “child guidance clinic.”

- *Lives Through Time* (Block, 1971): A detailed study of archival data of two samples of youth followed through adolescence and measured again in their late 30’s. The resulting work provides a detailed summary of continuity and change across major developmental milestones.

- *Dunedin Multidisciplinary Health and Development Study* (Dunedin Study; [http://dunedinstudy.otago.ac.nz](http://dunedinstudy.otago.ac.nz)): The Dunedin Study uses a modern prospective design, following a large group of individuals (N=1037) born in New Zealand between 1972-1973. Now in its fifth decade, the study has produced a wealth of findings relevant to clinical psychology.

- *Children in the Community* (CIC; [http://nyspi.org/childcom/](http://nyspi.org/childcom/)): Billed as a longitudinal study of mental health, the CIC has followed 800 children since 1975, assessing emotional and behavioral problems in the youth. Additional features include sub-studies of the offspring of the original sample and detailed life history interviews.

- *Pittsburgh Life History Studies* (PLHS; [http://www.lifehistorystudies.pitt.edu](http://www.lifehistorystudies.pitt.edu)): The PLHS includes the Pittsburgh Youth Study, started in 1987 to track the intellectual, social, and emotional development of inner-city boys, and the similarly tasked Pittsburgh Girls Study, started in 2000, focusing on inner-city girls. These studies have adopted an accelerated longitudinal design, which samples from cohorts that span 4-5 years to increase the ages covered by the studies.

- *Midlife in the United States* (MIDUS; [http://www.midus.wisc.edu](http://www.midus.wisc.edu)): The initial wave of MIDUS, collected in 1995-1996, was intended to provide broad measurement of behavioral, social, psychological, and health factors in a large nationally representative sample augmented by additional samples of interest (e.g., twins). There have since been two additional waves of
data, separated by approximately a decade each.

- **Minnesota Twin and Family Study (MTFS; [https://mctfr.psych.umn.edu](https://mctfr.psych.umn.edu))**: The MTFS was designed to measure genetic and biological influences on psychological functioning. Since 1987 the MTFS has followed large samples of twins longitudinally. The traditional emphasis on behavioral genetic designs of behavioral traits has been augmented with molecular genetic, magnetic resonance, and familial social behavior sub-studies.

- **Longitudinal Studies of Personality Disorder**: In the 1990’s three prospective naturalistic studies of personality disorder were started, each with diverse sampling strategies, but with a shared aim to test basic assumptions about this class of disorders. These include the Longitudinal Study of Personality Disorders (Lenzenweger, 2006), which followed high-risk and low-risk samples of university students throughout college; the Collaborative Longitudinal Personality Disorders Study (Skodol, 2005), which followed adult patients in four diagnostic groups over 12 years; and the McLean Study of Adult Development (Zanarini, Frankenberg, Hennen, Reich, & Silk, 2005), which followed a large sample of treatment seeking patients with Borderline Personality Disorder over 16 years and as of this writing is ongoing.

- **The Berlin Aging Study (Baltes & Mayer, 1999)**: A sample, initially of 516 individuals, aged 70-100+ followed since the early 1990’s, and measured on mental and physical health, as well as social and economic functioning. The sample size has decreased dramatically over time due to participant mortality, allowing for the study of this ultimate of outcomes.

- **Daily Measure of Affect (Diener & Larsen, 1984)**: This study is notable for its early use of EMA to sample affect, behavior, and situational characteristics twice daily form 42 undergraduates over the course of 6 weeks.

- **Daily Measurement of the Big-Five Traits (Borkenau & Ostendorf, 1998)**: A study of 22
participants who completed daily ratings of personality adjectives for 90 consecutive days. Notably these were then subjected to both idiographic (P-Technique factor analyses) and nomothetic factor analyses to compare individual personality structures to the structure of individual differences.

- **Traits as Density Distributions of States** (Fleeson, 2001): Sampling of personality trait relevant behavior several times per day over 2-3 weeks. As with Borkenau and Ostendorf (1998), this study is as remarkable for its analytic approach and conclusions as it is for its data collection strategy.

- **Momentary Assessment of Smoking Relapse** (Shiffman et al., 1996): Early use of portable computers to monitor smoking behavior in individuals who had recently quit, including initial lapse and behavior and psychological response.

**Major Accomplishments**

Here we summarize several of the major accomplishments of longitudinal designs. Similar to the landmark studies, what constitutes a major accomplishment depends, in large part, on one’s area of interest. Therefore, we focus on generalities that are likely to cut across areas of clinical psychology, even as we offer specific examples to provide context.

Among the major contributions of longitudinal designs is the ability to test basic assumptions about the stability and course of psychological constructs. To provide a concrete example of an area where longitudinal research has been transformative, we consider personality and its pathology. Going back to seminal writing at the end of the 19th century, personality was cast as immutable, at least beyond the age of 30 (James, 1950/1890). This assumption of unwavering stability was also naturally applied to personality disorders (American Psychiatric Association, 2013). Yet these assertions were based on scant data, and without the benefit of
systematic investigation they reflected little more than lay assumptions. Based on a variety of longitudinal studies, it was found that, while highly stable, personality shows normative developmental trends across the life span, from childhood to old age. Individuals differ in their stability, with some showing dramatic change, and others remarkable stability (Nesselroade, 1991). Perhaps not surprisingly, when the same types of investigations were applied to personality disorders they were found to be decidedly unstable, with individuals showing dramatic rates of remission over the relatively short periods of a year, significant rates of average decline when considering dimensional symptom counts, but also individual heterogeneity in these trajectories. Thus, the fundamental assumptions related to personality stability have themselves changed over time with the accumulating results of longitudinal studies.

In a related vein, personality was not only considered to be stable over the long-term, but also consistent across situations. In contrast, affect was long considered to highly variable, by definition, and was therefore frequently studied using methods that can capture moment-to-moment fluctuations. Eventually, however, methods capable of capturing variability in personality states were brought to bear on this question with startling results. For instance, moment-to-moment ratings of personality vary just as much as affect (Fleeson, 2001). Moreover, personality varied just as much within a person as it did between people. At the same time, weekly averages in personality rating were highly stable, suggesting that personality is highly stable over the short-term, in the aggregate. This shows how a concept can change with sampling using an appropriate time frame.

Although recent longitudinal research in personality is notable for the level of instability it has uncovered, other research has demonstrated that the psychology of the individual often shows powerful effects that transcend long periods across the life span. For instance, a variety
studies have converged on the finding that individual differences in behavioral regulation evident in early childhood predict important life outcomes (e.g., educational, occupational, relational attainment) through middle age. Additionally, studies in the area of health psychology have shown that basic dispositional features confer risks for the development of major physical illness and mortality that often exceed those of major risk factors like IQ and socioeconomic position. Importantly, variables measured at during one life period (e.g., childhood, early adulthood) that are predictive of later physical, emotional, and social functioning imply some mechanism that may serve as a target for intervention. Yet it is only by following individuals longitudinally that researchers are able to tease apart the competing yet oftentimes correlated pathways of risk and resilience. This, for example, has been one of the major goals of the landmark MIDUS study as it relates to midlife, and the Pittsburgh Life History studies as they related to disadvantaged youth.

Still other studies have incorporated advanced designs that allow for the adjudication between heritability and life experience. For instance incorporating biometrically informative design features, landmark longitudinal studies such as the Minnesota Twin and Family Study (MTFS) have illuminated the structure of etiologic processes, in terms of how they unfold. These insights have allowed better understanding of the broad etiologic underpinnings of major individuals differences, such as those of personality and psychopathology. MTFS has demonstrated, for example, that early disinhibitory personality features are genetically related to later substance use problems (Iacono & McGue, 2002), shedding light on how major forms of personality and psychopathology relate to one another. By controlling for genetic background variables, moreover, such studies also provide information on how environmental variables operate. Findings from MTFS, for example, have also shown that environmental variables
moderate the relationship between personality and psychopathology, showing that positive peer influences and attachment to parental and support institutions buffer against psychopathology among those at high risk (Iacono & McGue, 2002).

A consistent theme in this chapter has been that longitudinal designs can and do differ dramatically in the temporal distance between assessments and intensity of measurement. Intensive longitudinal designs have been gaining in popularity considerably since the 1990’s. In large part, this is because they offer incredible power to pose mechanistic questions about psychological processes on (or close to) the time scale at which they are hypothesized to function. Furthermore, by sampling from individuals in their environment greater accuracy of measurement and ecological validity is obtained. Accordingly, the development and refinement of the intensive longitudinal design has been a major accomplishment in this area of methodology.

To provide a concrete example of a research area that has benefitted immensely from the implementation of intensive longitudinal methods, we consider Saul Shiffman’s (2005) program of research on smoking quit and relapse. Events like smoking relapse following a quit attempt pose perplexing challenges to researchers, because they reflect a deviation from a process the individual is, arguably, motivated for and actively engaged in (i.e., abstinence). Along the way toward full relapse are often the initial lapses, which offer a point of intervention if only they can be understood, forecasted, and thwarted. Shiffman and his colleagues, over a series of studies using intensive longitudinal designs in naturalistic settings, have found that negative affect and stress both predict a smoking lapse. However, the precise timing was key, such that daily differences in stress and negative affect did not reliably predict lapses, but starting approximately 7hrs prior to a lapse, stress and negative affect demonstrate a linear trend toward the likelihood
of a lapse. Thus, it is not the background context of stress and affect, but rather a temporally proximal momentary increase that drives the lapse. Additional work found that daily self-efficacy was not predictive of lapses, but instead lapses contributed to significant drops in self-efficacy, as well as more nuanced contextual interactions between efficacy, affect, and lapses. By uncovering the nature and timing of these processes this type of research is more easily taken from bench to bedside with actionable recommendations.

**Future Directions**

The future of longitudinal designs appears very bright. Several key lines of progress are converging to provide a wealth of possibilities not previously available to clinical psychologists. These include rapidly developing technology for data capture across diverse modalities, cheap but powerful computing, and dissemination of advanced but user-friendly software. These in turn are motivating more ambitious longitudinal studies and increased sophistication in the questions being posed. They are providing additional potential tools for the practitioner interested in sampling patient behavior over time. At the same time, these technological advances are challenging psychologists to articulate, with much greater specificity, the precise nature of the processes and mechanisms they wish to measure. In this final section we pose several questions and forecast the next major steps in longitudinal design.

**Who will be sampled?**

We anticipate that much of the future longitudinal designs will be driven by standard sample selection procedures. We see two major departures from this structure that will impact research going forward. First, as there is increasing recognition that clinical constructs of interest are finely graded in the population (e.g., Markon, 2010; Wright et al., 2013), future longitudinal designs should increasingly shift from “group-based” design (e.g., Depressed vs.
Non-Depressed), and instead move toward designs that sample individuals spanning the range of phenomenological expression (i.e., dimensional clinical constructs). Second, as we all opt in to having more data passively collected about ourselves (e.g., via the Facebook, Google, activity monitors, Amazon.com), we anticipate that these public and private population-wide data will be increasingly mined to pose interesting questions about clinically relevant functioning (e.g., What are the electronically public behaviors that signal a suicide attempt or completion? What types of activity patterns predict a cognitive decline in advancing years?) making many more of us the participants of interest. In practical settings this may lead clinicians to find ways to also tap into and monitor this material over time.

**What variables will be collected?**

Moving forward, longitudinal designs will be increasingly multivariate and multimodal in their data collection. Studies that focus on one variable (e.g., depression symptoms) and/or a single data collection method (e.g., self-report, psychiatric interview) will be increasingly rare. Instead, with the now well established fact that diverse psychological, behavioral, and physiological systems all interpenetrate and reciprocally influence each other, the successful studies of the future will be purposefully designed to include a wide range of variables in an effort to provide greater coverage of the matrix of variables of interest to clinical psychologists. In the same vein, it is now widely accepted that no level of analysis (e.g., molecular, physiological, behavioral, self-report, etc.) can be fully privileged, and rather all tend to be incrementally informative. Therefore future studies will sample across levels of analysis.

**When will variables be sampled?**

With increasing variation in the possible timeframes examined in longitudinal designs, from intensive data collection involving multiple observations in a single day to long-term
designs over the course of years, it is critical to anticipate the timeframe over which changes of
interest might occur. For example, stress can be conceptualized as being relatively chronic or
acute, and depending on the process of interest, it might be more appropriate to collect
observations over a longer time period (in the case of chronic stressors, or major life events, for
example) or shorter time periods (in the case of acute stressors or daily hassles). Even in cases
where the relative timescale of a process might be understood in an abstract sense, it might be
challenging to decide exactly how to operationalize that timescale.

Consider studying short-term changes in internalizing problems and their association with
relationship conflict. Is it most appropriate to study phenomena over the course of weeks? Days?
Hours? In eras past, investigations and data-analysis were constrained to relatively limited study
designs of a cross-sectional or a handful of repeated measures. This limitation afforded theorists
the luxury of discussing processes in the abstract or without reference to precise time-scales and
temporal precedence. However, contemporary sampling and analytic methods should soon
compel greater specificity and clarity in articulation of psychological processes.

Accordingly, we expect to see designs sampling not only along one time-scale (e.g.,
across years, months, daily, etc.), but rather combining data collection at different temporal
levels of solution. For instance, using a longitudinal burst design to capture changes in micro-
processes as reflective of longer-term development (Sliwinski, 2008). Additionally, we
anticipate that scientific and practical interest in relatively infrequent events will invoke event
contingent sampling combined with bursts of short-term measurement. For instance, behavior
such as intimate partner violence, suicide attempts, and substance binges are chronic in some
individuals, but episodic and infrequent in others. To capture the full spectrum of the
phenomenon, and especially include those with a more episodic course, data collection may be
designed to have participant or patient initiated data collection right before, during, or after the
target event, which then triggers a series of automatic or investigator/clinician driven data
samplings in the period that follows. Repeated sampling in this fashion will allow for a better
understanding of the course of these behaviors and the mechanisms and processes associated
with them.

**Where will variables be sampled?**

The laboratory and consulting room will undoubtedly continue be important venues for
data collection in perpetuity due to the control it affords and advanced instrumentation it
supports. However, future designs will emphasize measuring variables of interest in daily life as
it is lived. Mobile, networked information devices have become an everyday part of life for
much of contemporary society, whether they are in the form of smartphones, tablets, activity
monitoring devices, or other tools. The ubiquity of such devices has created a flood of
possibilities for longitudinal studies. Whereas intensively longitudinal designs once relied
exclusively on dedicated devices provided by the researcher (and often still do), the future of
such data collection is arguably in smartphones and similar platforms, which allow for collection
of the data in a more naturalistic, less intrusive way. These devices can also be enlisted as
clinical tools with the practitioner providing the software to load onto the patient’s device.
Furthermore, the use of commonplace devices also masks the collection procedures that are
occurring in vivo, such that the participant or patient need not appear to be doing anything out of
the ordinary, thereby potentially ensuring greater compliance, privacy, and confidentiality. In
addition to these, ambulatory assessment of physiological variables (e.g., heart rate, blood
pressure) allows for the sampling of information usually considered the purview of laboratory
studies in the rich environment of participants' actual daily lives.
How will they analyzed?

Although this chapter is not intended to provide information on quantitative methodology, we would be remiss if we did not briefly mention how these new trends in data collection will be put to use. One direction that is gaining interest is the modeling of an individual’s data sampled intensively over time. Driving this interest is the recognition that extant psychiatric diagnoses are poorly suited to individual patients’ presentations, while at the same time practitioners require valid models for individual patients. Intensively sampled data allow for the application of complex statistical techniques that will effectively provide an individually tailored model for a patient that a practitioner can use to develop a personalized intervention (Van Os, Delespaul, Wigman, Myin-Germey's, & Wichers, 2013). We anticipate that analytic techniques capable of modeling complex dynamic processes (e.g., dynamical systems analysis), originally developed in other disciplines, will be imported and assimilated into psychological analysis of the individual. Good clinicians inherently work with a model for an individual, but a challenge facing research in this area of analysis is the integration of idiographic models across large samples of individuals. A further implication of this direction is that clinical training will likely require increased quantitative methodology training that goes beyond standard cross-sectional psychometrics and statistics, to incorporate these novel approaches.

As the timescales of longitudinal studies become more diverse, comprising scales of hours to weeks to years, and particularly when studies collect data on different temporal scales, it will be increasingly important to develop quantitative models that can accommodate changes at those various scales and integrate and relate changes at different timescales. For instance, a wide variety of classical models of change, have focused on trajectories of change that are somewhat static, in that an individual's change over time is modeled as a fixed trajectory, such as a linear or
It is possible to think of change itself as an input, however, as is reflected in more dynamic models of change (e.g., Boker, Neale, & Rausch, 2004) where the state at one timepoint relates to rates of change at other timepoints. One might even imagine change at one scale being modeled as a function of change at another; the rate of change of positive emotion over the course of a day, for example, might predict degree of change in relationships over a longer time period, such as a month or year, and that change in turn might predict daily changes in mood.

Research in the areas of substance abuse (Boker & Graham, 1998) and emotional regulation (Steele & Ferrer, 2011), for example, have demonstrated how modeling the effects of rate of change at different time points can be key to understanding dynamic processes. In this way, research into “change in change” and how to model it is likely to become increasingly critical in the future. Other aspects of change and stability over time, many of which are important to dynamic accounts, will also likely command more attention in future longitudinal modeling.

Longitudinal designs provide the clearest view into psychological mechanisms and are therefore an indispensable part of the clinical psychologist’s armamentarium. The power of the longitudinal design naturally comes at a cost, and a poorly designed data collection may not only be expensive but also offer little beyond a cross-sectional design. However, some of the most exciting research and practice trends in clinical psychology fall under the longitudinal design rubric. Methods that sample data intensively and repeatedly in naturalistic settings are providing a greater match between theoretical questions of interest, the data being collected, and the emerging parameters in statistical models. Due to the rapid increase in mobile computing, portable sensor arrays, and powerful analytic techniques, questions that previously seemed
purely the domain of theory are routinely being posed using hard data. The next several years will undoubtedly see the more widespread application of these techniques, as well as new advances in longitudinal methods.
References


Psychologist, 54, 5-20.


to age 9: Relations with gender, temperament, ethnicity, parenting, and rater.

*Developmental Psychology, 44*(3), 771-786.


psychology, 9, 151-176.


