Development and Psychopathology, **10** (1998), 395–426
Copyright © 1998 Cambridge University Press
Printed in the United States of America

The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations

JOHN B. WILLETT, JUDITH D. SINGER, AND NINA C. MARTIN
Harvard University

Abstract

The utility and flexibility of recent advances in statistical methods for the quantitative analysis of developmental data—in particular, the methods of individual growth modeling and survival analysis—are unquestioned by methodologists, but have yet to have a major impact on empirical research within the field of developmental psychopathology and elsewhere. In this paper, we show how these new methods provide developmental psychpathologists with powerful ways of answering their research questions about systematic changes over time in individual behavior and about the occurrence and timing of life events. In the first section, we present a descriptive overview of each method by illustrating the types of research questions that each method can address, introducing the statistical models, and commenting on methods of model fitting, estimation, and interpretation. In the following three sections, we offer six concrete recommendations for developmental psychopathologists hoping to use these methods. First, we recommend that when designing studies, investigators should increase the number of waves of data they collect and consider the use of accelerated longitudinal designs. Second, we recommend that when selecting measurement strategies, investigators should strive to collect equatable data prospectively on all time-varying measures and should never standardize their measures before analysis. Third, we recommend that when specifying statistical models, researchers should consider a variety of alternative specifications for the time predictor and should test for interactions among predictors, particularly interactions between substantive predictors and time. Our goal throughout is to show that these methods are essential tools for answering questions about life-span developmental processes in both normal and atypical populations and that their proper use will help developmental psychopathologists and others illuminate how important contextual variables contribute to various pathways of development.

Recent years have witnessed major advances in the statistical methods available for the quantitative analysis of longitudinal data. Descriptions of these advances—in particular, the methods of *individual growth modeling* and *survival analysis*—can be found throughout the technical literature and their strengths and generalizability are widely accepted among methodologists. Systematic inspection of issues of *Development and Psychopathology* over the last 9 years suggests, however, that—with a few notable exceptions—these innovations have yet to find their way into everyday empirical practice within the field of developmental psychopathology.

We believe that thoughtful application of these methods will help developmental psychopathologists better address research questions about the effects of context on development. Our goal in this paper, then, is to promote their proper use by demonstrating their utility, and by describing how developmental psychopathologists and others might think about structuring research projects so as to take fuller advantage of the methods'

The order of the first two authors was determined by randomization.

Address correspondence and reprint requests to: Judith D. Singer, Harvard University, Graduate School of Education, Larsen Hall, 7th Floor, Cambridge, MA 02138; E-mail: judith_singer@harvard.edu.

power. We begin with an introductory section that describes essential features of the two methods. Here, we give examples of the types of research questions that each can be used to address, we specify the underlying statistical models, we comment briefly on methods of model fitting and estimation, and we describe how statistical results can be translated into substantive findings. In the following three sections, we offer concrete recommendations for researchers contemplating use of the new methods—two about *research design,* two about *measurement,* and two about *statistical analysis.* First, we recommend that when designing studies, investigators should increase the number of waves of data they collect and consider the use of accelerated longitudinal designs. Second, we recommend that when selecting measurement strategies, investigators should strive to collect equatable data prospectively on all time-varying measures and should never standardize their measures before analysis. Third, we recommend that when specifying statistical models, researchers should consider a variety of alternative specifications for the time predictor and should test for interactions among predictors, particularly interactions between substantive predictors and time.

Statistical Models for the Study of Development and Psychopathology in Context

When investigators ask questions about human development, within both normal and atypical populations, they usually pose questions involving the passage of time. Broadly speaking, within this universe of research questions, we can distinguish between at least two important subclasses. One class of question focuses on the ways that individual attributes change over time. For example, in a study of the development of peer relations among elementary school children, Dodge, Pettit, and Bates (1994) ask how peer relations change as children mature and whether children who have been maltreated follow a different trajectory from those who have not. The other subclass focuses on the occurrence and timing of events. Researchers in this tradition ask whether individuals experience particular events or transitions, when these events occur, and what other variables predict variation in event occurrence and timing. In a study of juvenile delinquency, for example, Tremblay, Masse, Vitaro, and Dobkin (1995) ask (a) whether adolescent boys engage in delinquent behavior, (b) when these behaviors begin, and (c) whether age at onset is associated with friends' deviant behavior.

Addressing each of these two classes of question requires a different analytic strategy. The former requires methods for measuring and analyzing change—known variously as individual growth modeling (Rogosa, Brandt, & Zimowski, 1982; Willett, 1988), hierarchical linear modeling (Bryk & Raudenbush, 1992), random coefficient regression (Hedeker, Gibbons, & Flay, 1994), and multilevel modeling (Goldstein, 1995). The latter requires methods for analyzing the risk of event occurrence, known variously as survival analysis (Singer & Willett, 1991, 1993; Willett & Singer, 1993, 1995), event history analysis (Allison, 1984), and hazard modeling (Yamaguchi, 1991). Below, we outline briefly the salient features of each.

Measuring and Modeling Individual Change Within Context

When people acquire new skills, when they learn something new, when their attitudes and interests develop, they change in fundamental ways. Despite its importance, much controversy has surrounded the measurement of change (Rogosa et al., 1982; Willett, 1988, 1994). In the past, influential methodologists convinced themselves, and everyone else, that it was not possible to measure change well. Their widely publicized conclusions were rooted in a simple misconception—that individual change should be viewed as an increment—the difference between "before" and "after."[1]

[1] For a critical discussion of classical methods for the measurement of change, see Willett (1995, 1988), Rogosa and Willett (1985), and Rogosa, Brandt, and Zimowski (1982).

Methodologists now know that this perception is mistaken. Individual change takes place continuously over time, and comparison of each person's "before" and "after" status is not the most subtle, nor the most effective, way to reveal the features of that trajectory. To measure individual change well, a truly longitudinal perspective must be adopted—a sample of people must be followed over time allowing the researcher to collect multiple waves of data at sensibly spaced intervals.[2]

We illustrate the ideas behind individual growth modeling using data on the delinquent behavior of 124 adolescents who participated in the 1988, 1990, and 1992 administrations of the *Children of the National Longitudinal Survey of Youth* (NLSY).[3] In the left-hand panel of Figure 1, we display delinquent behavior scores for one of these respondents, a boy (ID 994001). In the panel, we plot his observed score on the vertical axis versus his age (here, 11, 13, and 15 years). Notice the trend in his empirical growth record—the observed scores increase with age, suggesting that he is engaging in greater amounts of delinquent behavior as he grows older.

Individual changes over time like these can be represented by an *individual growth model* that describes the temporal dependence of individual status on time. For these data, we might hypothesize that the delinquent behavior (DELBEH) exhibited by adolescent j on occasion i can be expressed as a linear function of AGE,

$$\mathrm{DELBEH}_{ij} = \{\pi_{0j} + \pi_{1j}(\mathrm{AGE}_{ij} - 11)\} + \varepsilon_{ij}, \quad (1)$$

where we have bracketed structural component of the model, representing the dependence of true delinquent behavior on time, to separate it from the random error, ε_{ij}, that accrues on each occasion of measurement. Equation 1 is often referred to as the "within-person" or "level-1" individual growth model. The structural part of the level-1 model contains unknown constants referred to as individual growth parameters, whose values determine the trajectory of true individual change over time. Equation 1 contains two such parameters: π_{0j} and π_{1j}. If an appropriate model has been selected to represent individual growth, these parameters represent key features of the true growth trajectory for person j. In this case, where the growth is hypothesized to be linear, π_{0j} represents the adolescent's true level of delinquent behavior at age 11 years and π_{1j} represents his or her true rate of change in delinquent behavior over time. If π_{1j} is positive, then child j's true level of delinquent behavior increases with time; if it is negative then it decreases. We have fit this model to adolescent 994001's data using ordinary least squares (OLS) regression and superimposed the fitted line on the left-hand panel of Figure 1. Notice that the estimated slope is positive (+2.0) indicating that, for this boy, delinquent behavior tends to increase during adolescence.

One important feature of the level-1 growth model is that the *researcher* controls the substantive interpretation of the intercept parameter, π_{0j}. By subtracting 11 from the adolescent's age before multiplying by the individual slope parameter (as in Equation 1), we have "recentered" the origin of the time axis to age 11 years. Recentering provides the individual intercept parameter with an interpretation that is substantively interesting in the context of this study—it represents true delinquent behavior on entry into the study at age 11 years. In the case of adolescent 994001, the OLS estimate of his initial level of delin-

[2] Note that the methods of individual growth modeling are only applicable if it truly makes sense to measure change in the attribute of interest. At the very least, the attribute must be a continuous variable, must be equatable over occasions of measurement, and must remain construct valid for the period of observation.

[3] Delinquent behavior was measured using nine items drawn from the NLSY. These items asked how many times, in the last year, did the adolescent stay out later than the parent said, stay out without parental permission, have to bring the parents to school, hurt someone enough for them to need a doctor, lie about something important, steal from a store, damage school property, get drunk, or skip school without permission. Respondents rated each item on a 4-point scale (0 = *never*, 1 = *once*, 2 = *twice*, 3 = *more than twice*). Individual responses were summed across the 9 items, providing an observed delinquent behavior score that could range from 0 to 27.

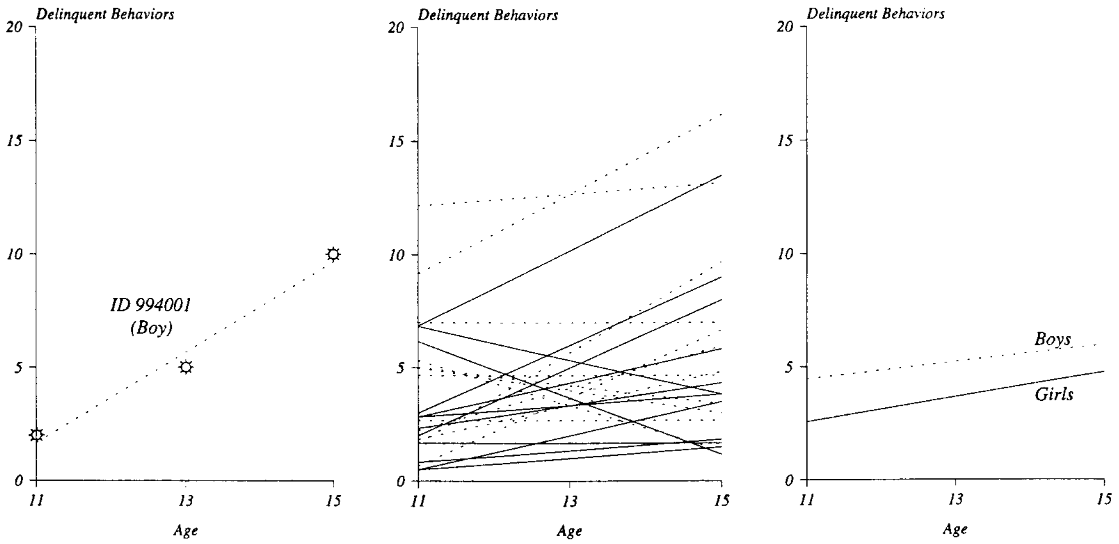

Figure 1. What happens when you fit individual growth models. Left panel: Observed scores and an empirical growth trajectory for a single case; center panel: a sample of OLS-fitted growth trajectories for a sample of 25 cases, coded by the gender of the respondent (boys are dashed; girls are solid); right panel: fitted individual growth trajectories for boys and girls corresponding to the model in Equations 1 and 2.

quent behavior is 1.67. As we describe in a later section on analytic recommendations, many alternative parameterizations of age are possible, each giving rise to a different interpretation of the intercept.

Just as we are not limited to a particular definition of the intercept, we are also not limited to a linear individual growth summary. Many other possible mathematical functions are available—both those that depend linearly on time, and those that do not. Choice of an appropriately shaped trajectory to represent true individual change is an important first step in any analysis. Ideally, theory will guide the rational choice of trajectory so that subsequent analyses have meaningful interpretations. Often, however, the mechanisms governing the change process are poorly understood and a linear or a quadratic curve is used to approximate the trajectory. Also, in much research in psychology and psychopathology, only a restricted portion of the life span is observed and few waves of data are collected; thus, the selected trajectory must be mathematically simple. Accordingly, the trend used to summarize individual change over time is often a linear function of time, as it is here. (Other possibilities will be explored later in the paper.)

A key assumption of individual growth modeling is that the trajectory for each person in the population has the same functional form—in this case, linear—but that different individuals may have different values of the individual growth parameters. Adolescents in this example may differ in their intercepts (some adolescents may display little delinquent behavior at age 11 years, some may display a lot) and in their slopes (the delinquent behaviors of some adolescents may change rapidly with age, while others may display behaviors that are relatively stable or even decline as time passes). Such heterogeneity can be seen in the center panel of Figure 1, where we display the OLS-fitted individual growth trajectories for 25 adolescents selected at random from the larger group of 124.

Notice that we have coded the trajectories by the gender of the adolescent—dashed lines for boys, solid lines for girls. Plots like these allow us to investigate whether individual growth trajectories differ from person to person and if the interindividual variation is systematically related to various contextual variables, such as characteristics of the individual, his or her family, or his or her community. Questions like these—about the correlates and predictors of change—naturally translate into questions about relationships between the individual growth parameters and variables representing individual (and group) characteristics. Inspecting the center panel of Figure 1, for example, we might ask whether boys and girls differ in either their delinquent behavior at age 11 years (represented by the individual intercepts) or the rate at which delinquent behavior changes with age (represented by the individual slopes). If we detect systematic interindividual variation in change, we know that children with different characteristics—for example, gender, family environment, treatment conditions—grow in different ways. Questions such as these provide an important window into the effects of context on development by allowing researchers to determine how individuals from diverse backgrounds may develop in ways similar to or dissimilar from one another. In this way, individual growth modeling may be said to be consistent with the "person-oriented level of analysis of a differential pathways approach" to developmental psychopathology advocated by Cicchetti and Rogosch (1996, p. 598), such that one might examine how a diverse set of contextual variables may lead to common outcomes among some individuals, or, conversely, how similar contexts may result in dissimilar outcomes among others.

Analytically, we specify a second statistical model—often called the "between-person" or "level-2" model—to represent interindividual differences (Bryk & Raudenbush, 1987; Rogosa & Willett, 1985). In the level-2 model, we express the individual growth parameters as a function of the selected characteristics. For example, to examine whether individual growth trajectories differ for boys and girls, we would posit the following pair of simultaneous level-2 models,

$$\begin{aligned} \pi_{0j} &= \beta_{00} + \beta_{01}\text{FEMALE}_j + u_{0j}, \\ \pi_{1j} &= \beta_{10} + \beta_{11}\text{FEMALE}_j + u_{1j}, \end{aligned} \qquad (2)$$

where the dichotomous predictor FEMALE_j indicates whether adolescent j is a girl and the level-2 residuals, u_{0j} and u_{1j}, represent those portions of the individual growth parameters that are "unexplained" by the selected predictor of change. The β coefficients summarize the population relationship between the individual growth parameters and the selected characteristics. They can be interpreted in much the same way as regular regression coefficients. For instance, if the level of delinquent behavior of girls at age 11 years happens to be higher than that of boys (i.e., if they have larger values of π_{0j}, on average) then β_{01} will be positive (since FEMALE = 1 for girls). If boys have higher *rates of change* in delinquent behavior (i.e., if they have larger values of π_{1j}, on average), then β_{11} will be positive (see below for estimates).

Researchers modeling change can fit the statistical models in Equations 1 and 2 to data, allowing estimation and subsequent interpretation of parameters. A variety of methods are available for fitting models and estimating parameters. Some methods are very straightforward and can easily be implemented on popular commercially available statistical computer packages; others are more sophisticated and require dedicated software.

The simplest approach is strictly exploratory, as we have already begun to demonstrate in Figure 1. Here, the level-1 individual growth model is fitted separately for each person in the data set by OLS regression. These "person-by-person" analyses provide individual growth parameter estimates for each person that can be collected together to become dependent variables in subsequent, and separate, between-person data analyses. For instance, in the case of Equation 1, we can first obtain individual intercept and slope estimates to represent delinquent behavior at age 11 and the rate of change in delinquent behavior by regressing observed delinquent behavior on age (minus 11 years, see Equation 1) for each person in the sample. These estimates can then be collected together and regressed directly on FEMALE, or other contextual predictors such as family structure, socioeconomic status, or neighborhood crime rate, in follow-up level-2 regression analysis.

This exploratory approach can be improved by accounting for interindividual differences in the precision of the growth parameter estimates. Due to idiosyncracies of measurement, some people may have empirical growth records whose entries are smoothly ordered and for whom the growth data fall very close to the underlying true trajectory. Other people may have more erratic growth records with their data points scattered widely from the underlying true trajectory. These differences in scatter affect the precision (the standard errors) with which the level-1 individual growth parameters are estimated. Those with smooth and systematic growth records will have more precise estimates of intercept and slope (that is, the parameter estimates will have small standard errors); those with erratic and scattered observed growth records will have less precise estimates. Level-2 analyses of the relationships between the estimated individual growth parameters and the predictors of change can be improved (made asymptotically efficient) if between-person variation in the precision of the first-round growth parameter estimates is taken into account (Willett, 1988).

These ideas are behind much of the dedicated computer software now available for fitting the level-1 and level-2 statistical models simultaneously. Kreft, de Leeuw, and Kim (1994) provide a comprehensive review of several of the programs that were available in the early 1990s. An exciting new development is the availability of routines for fitting these models in the major statistical packages. SAS now includes a dedicated procedure—PROC MIXED—that can be used to fit these models (see Singer, in press) as does STATA (XTREG). When data collection has been time structured—data are available on all subjects at the same ages—individual growth models can also be fit using the methods of covariance structure analysis (see Willett & Sayer, 1994).

We used SAS PROC MIXED to simultaneously fit the models in Equations 1 and 2 to our illustrative data. We present the results of fitting these models in the right-hand panel of Figure 1, which presents fitted growth trajectories for boys and girls. Interpreting the actual parameter estimates, we find that $\hat{\beta}_{00} = 5.2$, indicating that the average 11-year-old boy has a score of just over 5 on the delinquent behavior scale; (b) $\hat{\beta}_{01} = -1.55$, indicating that at age 11 years, the average girl scores about one and half points less than the average boy; (c) $\hat{\beta}_{10} = .38$ indicating that after age 11 years, the average boy grows just under four tenths of a point per year; and (d) $\hat{\beta}_{11} = .17$ indicating that the average annual growth rate for girls is .17 points higher than the average annual growth rate (.38) for boys.[4]

Individual growth modeling offers empirical researchers many advantages. The method can accommodate any number of waves of data, the occasions of measurement need not be equally spaced, and different participants can have different data-collection schedules. Essentially, then, not only is the method flexible enough for almost any empirical setting, but also the precision (and the reliability) with which change can be measured is under the direct control of the investigator via the manipulation of research design. As we describe later, individual change can be represented by a variety of substantively interesting trajectories, including straight-line, curvilinear, or even discontinuous functions. Finally, not only can multiple predictors of change (e.g., predictors that represent the context in which individuals develop) be included in the analysis, but simultaneous change across multiple domains (e.g., change in cognitive functioning and change in self-esteem) can be investigated simultaneously.

Measuring and Modeling the Risk of Event Occurrence in Context

A second class of question posed in developmental research asks "whether" and, if so, "when" particular events occur. In a recent book on stress and adversity across the life course (Gotlib & Wheaton, 1997), for example, researchers interested in the sequelae of trauma asked a variety of questions including *whether* an individual ever experiences depression and, if so, *when* onset first occurs (Wheaton, Roszell, & Hall, pp. 50–72);

[4] To plot the fitted growth trajectories, we substituted estimates of the four level-2 β coefficients into Equation 2 to generate estimates of individual intercept and slope for the average boy and girl. These estimated individual growth parameters were then used to generate the required trajectories.

whether and when street children returned to their homes (Hagan & McCarthy, pp. 73–90); whether and when high school graduates got married and began a family (Gore, Aseltine, Colten, & Lin, pp. 197–214); and, whether and when young children made the transition between adult supervised care and self-care (Belle, Norell, & Lewis, pp. 159–178).

Familiar statistical techniques, such as multiple regression and analysis of variance, are ill-suited for addressing such questions because they cannot handle situations in which the value of the outcome—in this case, whether and when an event occurs—is unknown for some people under study. Yet when event occurrence is studied, such an information shortfall is almost inevitable. No matter how long data are collected, some members of the sample will not experience the target event during data collection—some people will not get depressed, some street children will not return home, some high school graduates will not begin a family. We say that such observations are *censored,* and censoring creates an analytic dilemma. Although the researcher does, in fact, know something about individuals with censored event times—that is, if they do experience the event, it must be after data collection ends—this knowledge is imprecise. The dilemma is how to analyze data simultaneously from both censored and noncensored cases, because the censored members form a key group—they are often the ones least likely to experience the event.

The methods known variously as survival analysis, event history analysis, or hazard modeling provide this egalitarian level of inclusion. To use them, the researcher must record, from a predefined starting time, how long it takes each person in the sample to experience the target event. Typically, the researcher follows sampled individuals (either prospectively and periodically, or by retrospective event history reconstruction) and records whether and, if so, when the event occurs. All who experience the event during observation are assigned explicit event times. Those who do not experience the event during observation are noted as censored and the length of time that they went without experiencing the event is recorded. Subsequently, their "censored" lifetimes enter into the data analyses in a meaningful way.

Some researchers record event occurrence very precisely. When studying the relationship between childhood adversity and death, for example, Friedman, Tucker, Schwartz, and Tomlinson–Keasey (1995) used public records to determine the precise time (year, month, day) of death. We refer to such precise records of event occurrence as *continuous-time* data. More commonly, however, researchers record only that the event occurred within some finite time interval. A researcher might know, for example, the *year* that a person first experienced depressive symptoms or the *grade* that a child switched from adult-supervised care to self-care. We call data such as these *discrete-time* data. Because discrete-time data are so common in developmental studies, we focus on methods for these data in this paper, known as discrete-time survival analysis.

When examining the occurrence of an event such as "experiencing an initial episode of depression" for a random sample of individuals, we begin by investigating the pattern of event occurrence over time. We ask, for example, when are individuals most likely first to experience a depressive episode—during childhood, their teens, or their 20s, 30s, or 40s? When we pose such questions, we are implicitly asking about variation in the risk of event occurrence across time periods. Knowing how the risk of experiencing an event fluctuates over time answers both the whether and the when questions posed.

But how can the risk of event occurrence be summarized, especially when some of the sampled people have censored event times? In discrete-time survival analysis, the fundamental quantity representing the risk of event occurrence in each time period is called the *hazard probability.* Its computation in the sample is straightforward. In each time period, one must identify the risk set—the pool of people who are at risk of experiencing the event in this period (i.e., those who have reached this time period without experiencing the event)—and compute the proportion of this group that experiences the event during the period. Notice that this definition is inherently condi-

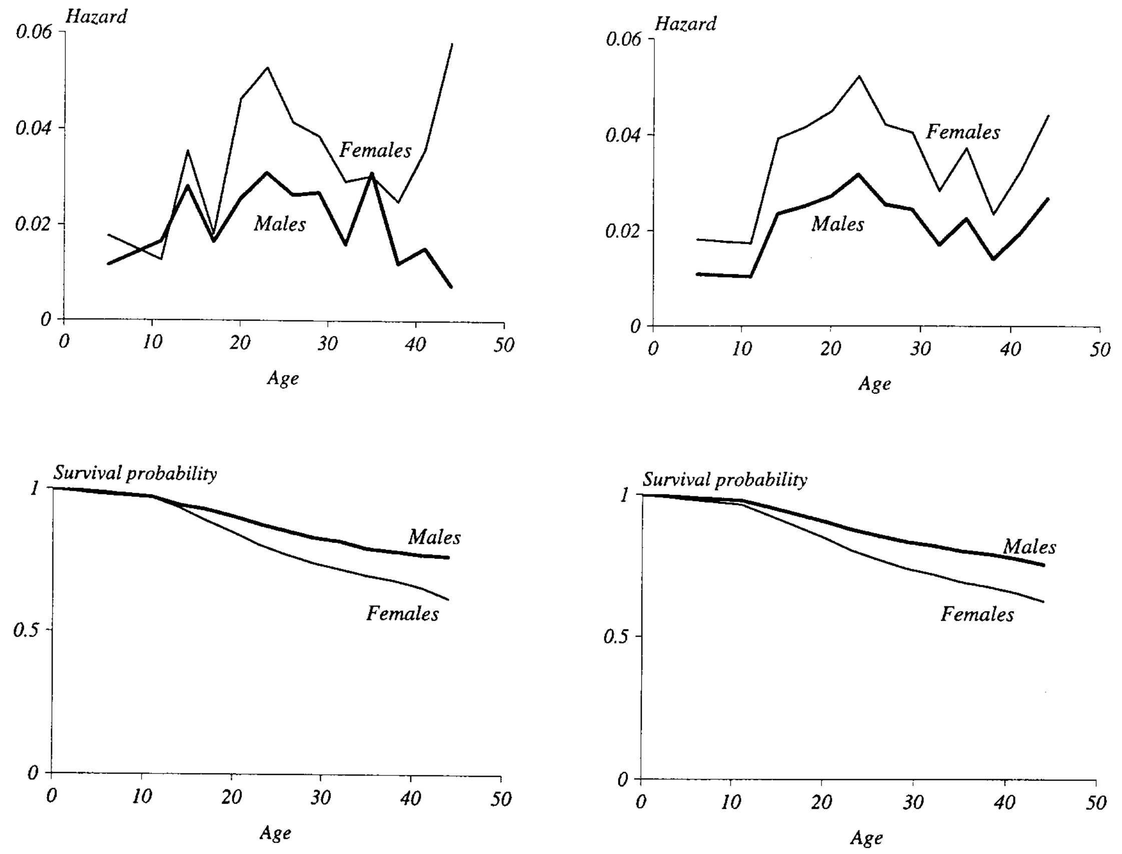

Figure 2. What happens when you fit hazard models. Hazard and survivor functions describing age at first onset of depression by gender. The left panel presents *sample* functions; the right panel presents *fitted* functions.

tional: once a person experiences the event (or is censored) in one time period, he or she is no longer be a member of the risk set in any future period. A plot of the set of hazard probabilities against time yields the *hazard function,* a chronologically ordered summary of the risk of event occurrence.

In the top left-hand panel of Figure 2, we present sample hazard functions estimated from retrospective data on 1,393 Canadian adults who were asked whether and, if so, when they first experienced a depressive episode (Wheaton et al., 1997). These functions describe the risk of initially experiencing a depressive episode in each of 13 successive time periods (age 9 or younger, 10–12, 13–15, 16–18, . . . , 40–42, and ages 43 years and older). Inspection of the sample hazard functions helps pinpoint when events are likely to occur—we see that for both males and females, the risk of experiencing an initial episode of depression is low in childhood, increases during adolescence, and then peaks in the early twenties. After this point, the risk of initial onset of depression, among those who have not yet had a depressive episode, is much lower. By the early forties, the risk declines to preadolescent levels for men but rises again for women. Beyond this overall pattern, notice that in all but two time periods, a sex differential exists—women seem to be at greater risk of experiencing a depressive episode than men.

The "conditionality" inherent in the definition of hazard is critical because it leads the hazard probability to deal evenhandedly with censoring by ensuring that all individuals remains in the risk set until the last time period that they are eligible to experience the event (at which point they are either censored or they experience the target event). For example, the hazard probability for initial onset of depression during the age period 31–33 years is estimated conditionally using data from all those individuals (852 of the initial sample of 1,393) who were at least age 31 when data were collected, but who *had not yet had a depressive episode during any earlier time period.* Individuals who were not yet in their early thirties ($n = 227$) or who had already experienced a depressive episode ($n = 314$) were no longer at risk and were excluded from the computation of hazard in this time period and all subsequent time periods.

In addition to using the hazard function to display the risk of event occurrence over time, the period-by-period risks can be cumulated to display the proportion of a sample that "survive" through each time period without experiencing the event. This proportion is called the *survival probability,* and a *survivor function* is a plot of this proportion against time (for computational details, see Willett & Singer, 1993). In the bottom left-hand panel of Figure 2, we display sample survivor functions for the men and women in our example. These functions present the proportion of adults who "survived"—that is, did not experience an initial depressive episode—through each successive time period. Notice that the curves are high in the beginning—at birth, all individuals are "surviving," as no one has experienced a depressive episode and thus the survival probabilities are 1.00. Over time, as individuals begin to experience initial depressive episodes, the survivor functions decline. Because most adults in this sample never experience a depressive episode at any time in their lives, the curves do not reach zero, but end at .77 for men and .62 for women.

Sample hazard and survivor functions describe whether and when individuals are likely to experience a target event. They can also be used to answer questions about group differences that represent the differing contexts in which individuals develop. Such contextual variables and the associated research questions that might be addressed could include, for example, family size—are individuals from larger families less likely to experience a depressive episode than individuals from smaller families?; child maltreatment—are maltreated children more likely than nonmaltreated children to repeat a grade in school?; or parental divorce—are children of divorced parents more likely than children of intact families to undergo a divorce themselves? Implicitly, each of these examples uses individual contextual characteristics—family size, child maltreatment, and parental divorce—to predict the risk of event occurrence. When we contrast the pairs of sample hazard and survivor functions displayed on the left-hand side of Figure 2, we are implicitly treating gender as a predictor of risk of first onset of depression. But such exploratory comparisons are limited because, using sample plots, it is difficult to examine the effects of continuous predictors, to examine the effects of several predictors simultaneously, to explore statistical interactions among predictors, and to make inferences about the population from which the sample was drawn. These more complex analytic goals are achieved by postulating and fitting statistical models of the hazard function and by conducting tests on the parameters of these models.

Statistical models of hazard express hypothesized population relationships between entire hazard profiles and predictors. To motivate our representation of this idea, examine the two sample hazard functions in the top left panel of Figure 1 and imagine that we have created a dummy variable, FEMALE, taking on values of 0 for males, 1 for females. In this formulation, visualize the entire hazard function as the conceptual "outcome" and the dummy variable FEMALE as the potential "predictor." How should we characterize the relationship between outcome and predictor? Ignoring differences in the shapes of the profiles for the moment, notice that when FEMALE = 1, the sample hazard function is generally "higher" relative to its location when FEMALE = 0, indicating that in virtually every time period, women are more likely to experience an initial depressive episode. So conceptually, at least, the effect of the predictor FEMALE seems to be to "shift" one sample hazard profile vertically relative to the other. A population hazard model formalizes this conceptualization by ascribing vertical displacement in the population hazard profile to variation in the predictors.

The complication, of course, is that the discrete-time hazard profile is no ordinary continuous outcome. It is a set of conditional probabilities, each bounded by 0 and 1. Statisticians who model a bounded outcome like this as a function of predictors generally do not use a linear function to express the relationship. Instead, they use a nonlinear link function that has the net effect of transforming the outcome so that it is unbounded, in order to prevent fitted values from falling outside the permissible range (in this case, between 0 and 1). When the outcome is a probability, as it is here, the logit link function is popular (Collett, 1991). If p represents a probability, then logit (p) is the natural logarithm of $p/(1-p)$ and, in the case of these data, can be interpreted as *the log-odds of initial onset of depression.*

Letting $h_j(t_i)$ represent the population hazard profile—that is, a list of population conditional probabilities for person j at discrete times, t_i, a suitable statistical model relating the logit transform of hazard to values of the predictor FEMALE is

$$\text{logit } h_j(t_i) = \beta_0(t) + \beta_1\text{FEMALE}_j, \qquad (3)$$

where parameter $\beta_0(t)$ is known as the *baseline logit-hazard profile.* It represents the value of the outcome (the entire logit-hazard profile) when the value of the predictor FEMALE is 0 (i.e., it specifies the profile for men). Notice that we write the baseline as $\beta_0(t)$, a function of time, and not as β_0, a single term unrelated to time (as in regression analysis), because the outcome (logit $h(t)$) is an entire temporal profile. The discrete-time hazard model in (3) specifies that differences in the value of the predictor "shift" the baseline logit-hazard profile up or down. The magnitude of the "slope" parameter β_1 represents the vertical shift in logit-hazard associated with a one unit difference in the predictor. Because the predictor here is dichotomous, β_1 captures the differential risk of onset (measured in the logit hazard scale) for women in comparison to men.

Model fitting, parameter estimation, and statistical inference for discrete-time hazard models are easily achieved using standard software for logistic regression (for a technical discussion, see Singer & Willett, 1993; for a hands-on applied discussion, see Willett & Singer, 1993). Without delving into details, suffice it to say that once a discrete-time hazard model has been fit, its parameters can be reported along with standard errors and goodness-of-fit statistics in much the same way that the results of regular regression analyses are reported. And, just as fitted lines can be used to illustrate the influence of important predictors in the context of multiple regression, so, too, can fitted hazard functions (and survivor functions) be displayed for prototypical people—those who share substantively important values of selected predictors.

We illustrate the results of this process in the right-hand panel of Figure 2 which presents fitted hazard and survivor functions for the model presented in Equation 3. Comparing the right and left panels, notice that the fitted plots on the right side are far smoother without the crossing and zig-zagging characteristic of the sample plots on the left side. This smoothness results from the constraints inherent in the population hazard model stipulated in Equation 3, which forces the vertical separation between the two hazard functions to be identical (in logit-hazard scale) in every time period. Just as we do not expect a fitted regression line to go through every data point in a scatterplot, we do not expect a fitted hazard function in survival analysis to match every sample value of hazard since the discrepancies between the sample and fitted plots presented in Figure 2 may be nothing more than sampling variation.

What have we learned by fitting this statistical model to these data? First, we reveal a more clearly articulated profile of risk over time by pooling information across individuals and by asking questions about the population from which these data derive. Here, our analyses concur with the findings of other researchers who have studied the initial onset of depressive disorders (e.g., Sorenson, Rutter, & Annenschel, 1991): the risk of onset is relatively low in childhood, rises steadily through adolescence, reaches a peak in the early twenties, at which point it declines, falling not back to zero, but to moderate levels that never quite reach the peak risks of early adulthood. Second, we can quantify the increased risk of initially becoming depressed among women in comparison to men, and we can conduct a hypothesis test of whether this gender differential may be a result of sampling variation. Our analyses yield a parameter estimate of 0.52 for β_1, indicating that the vertical separation, in the logit-hazard scale, between the profiles of risk for men and women is 0.52. Conducting the appropriate hypothesis test, we obtain a χ^2 test statistic of 23.20 ($df = 1$, $p < .0001$) and can therefore reject the null hypothesis that the predictor FEMALES has no effect on the population hazard profile (i.e., we reject the null hypothesis that H_0: $\beta_1 = 0$). Because few researchers possess an intuitive understanding of the logit-hazard scale, we recommend using the standard data-analytic practice of antilogging the coefficient in order to interpret it in terms of odds and odds-ratios (Hosmer & Lemeshow, 1989). Antilogging .52, the estimated odds of experiencing an initial depressive episode in any given time period are 1.67 times higher for women compared to men (again confirming other investigators' findings that women typically display higher levels of internalizing behaviors, such as depression, than do men; Kandel & Davies, 1986; Nolen–Hoeksema, 1990; Petersen, Sarigiani, & Kennedy, 1991).

The fitting of discrete-time hazard models provides a flexible approach to investigating predictors of event occurrence that appropriately includes data from both censored and non-censored individuals. Although hazard models may appear unusual, they actually resemble familiar multiple linear and logistic regression models. Like these familiar models, hazard models can incorporate several predictors simultaneously, permitting the examination of the effect of one predictor while controlling statistically for the effects of others. In this way, then, developmental psychopathologists might, for example, study the effect of maternal depression on the prediction of the onset of disruptive behavior problems in children while controlling for the effect of family socioeconomic status. Given that low socioeconomic status is often associated with multiple risk factors, such as maternal depression (Shaw, Owens, Vondra, Keenan, & Winslow, 1996), controlling for SES within a discrete-time hazard model can allow investigators to examine the effect of a contextual variable such as maternal depression over and above the effect of the context of family poverty. Similarly, we can examine the *synergistic* effect of several contextual predictors by including statistical interactions among them. Accordingly, then, one might study how the effect of maternal depression on the onset of childhood disruptive behaviors differs in families *below the poverty line* versus those *above it,* affording yet another view of the importance of contextual variables in the prediction of the development of maladaptive behavior over time. Such a view of the interactive nature of determinants of developmental pathways is consistent with the conceptualization of developmental psychopathology, espoused by many researchers, as a "developmental process in which the individual's adaptive functioning at any point in time is the product of multiple, interacting factors, including contextual and organismic variables" (Walker, Neumann, Baum, Davis, DiForio, & Bergman, 1996, p. 655). As is the case with individual growth modeling, then, discrete-time hazard models allow for the investigation of those variables that characterize the specific context in which development occurs and thus offer investigators valuable tools for the study of developmental psychopathology in context.

As alluded to above, one appealing feature of hazard models is that we can include predictors whose values *vary with time.* Unlike time-invariant predictors, such as sex or race, time-varying predictors describe contextual characteristics that may fluctuate with time, such as an individual's marital status, income, level of depression, or exposure to life stress. For clarity, when specifying statistical models that include time-varying predictors, we include a parenthetical *t* in the variable name to distinguish time-varying predictors from their time-invariant cousins. We believe that the inclusion of time-varying predictors in hazard models represents an exciting opportunity for two reasons. *First,* when investigating development, researchers often study behavior across extended periods of time and it is natural for the values of substantively important predictors to vary. In the investigation of schizophrenia, for example, studies have shown that exposure to life stress, such as parental maltreatment, is related to the expression of the genetic predisposition (i.e., the congenital diathesis) for schizophrenia (Walker et al., 1996). Certainly, life stress is not a "static phenomenon," such as gender, but one whose level, and thus effect on an outcome such as the expression of schizophrenia, changes over time, often as a result of other time-varying predictors, such as family socioeconomic status. This consideration of variables changing in concert with one another brings to mind a *second* reason that the inclusion of time-varying predictors in hazard models represents an exciting research opportunity: research questions about developmental processes of adaptation and maladaptation often focus on the co-occurrence of several different events. Developmental psychopathologists may ask, for example, whether the occurrence of one stressful event, such as parental divorce, predicts the occurrence of another stressful event, such as the onset of depression. Such questions can be answered simply by coding the precipitating event as a time-varying predictor.

We illustrate the use of time-varying predictors by adding the dummy variable $\text{PARDIV}_j(t_i)$, which indicates whether individual j's parents had divorced by time t_i (0 = *not yet divorced;* 1 = *divorced*), as a predictor to our previous logit-hazard model, Equation 3[5]:

$$\text{logit } h_j(t_i) = \beta_0(t) + \beta_1\text{FEMALE}_j + \beta_2\text{PARDIV}_j(t_i). \quad (4)$$

In Equation 4, the values of predictor PARDIV(t) vary over time (beginning at 0 among intact families and switching to 1 if, and when, the individual's parents divorce). The model stipulates, however, that the effect of parental divorce on the risk of onset is constant over time, represented by the single parameter β_2. Here, we estimate β_2 to be 0.34, indicating that the odds that a child of divorced parents will become depressed are 1.41 ($=e^{0.34}$) times higher than the corresponding odds for a child of nondivorced parents. (Later in the paper, we will show how to relax the assumption that the effect of a predictor is constant across the life span.)

Figure 3 presents the results of fitting the model in Equation 4. Comparison of the four prototypical hazard functions illustrates the large and statistically significant effects of the two predictors: Women are at greater risk of experiencing depression as are individuals whose parents divorced. Because PARDIV(t) is a time-varying predictor, however, these fitted functions should not be interpreted in exactly the same way as the fitted plots in Figure 2. Focus first on the bottom fitted hazard profile, which depicts the risk of experiencing a depressive episode among men whose parents never divorced. This profile is the lowest of the four fitted hazard profiles because this group is at lowest risk of experiencing a depressive disorder. Now consider the profile that would result if a boy's (or man's) parents divorce. While the parents were married, the boy's risk profile would still be represented by the lowest of the four hazard functions. When they divorce, however, the later portion of this boy's risk profile (after the divorce) would be described by the other fitted hazard profile for males, which is substantially higher, capturing the increased risk of depression among males whose parents had divorced.

As with growth modeling, the advent of hazard modeling offers much to developmental psychopathologists and others who seek to study development in context. Not only can the occurrence and timing of events be investigated within a coherent framework, but the ease with which time-varying predictors can be incorporated offers a unique analytic opportunity. Given that many contextual predictors fluctuate naturally with time (e.g., family and social structure, employment, opportunities for emotional fulfillment, and exposure to extreme life stress), hazard modeling allows investigators to study how various

[5] Additional analysis confirmed that no statistical interaction existed between these main effects—that is, the effect of parental divorce on risk was identical for men and women.

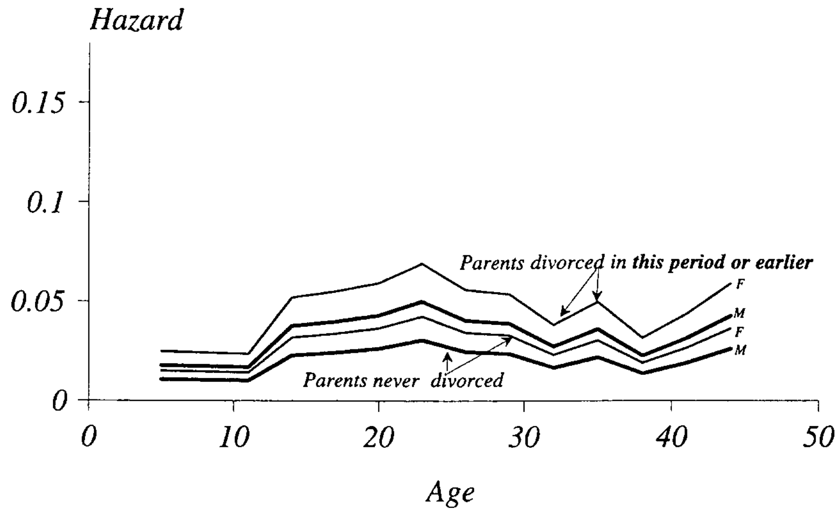

Figure 3. Including a time-varying predictor in hazard models. Fitted hazard functions describing age at first onset of depression by gender and whether the respondent's parents had divorced.

life contexts eventuate in a variety of developmental pathways, allowing for the consideration both of how *different life contexts* may lead to *similar outcomes* (a process described by Cicchetti, 1990 as "equifinality") and how *similar life contexts* may lead to a variety of *different developmental outcomes* (Cicchetti's, 1990, principle of "multifinality"). In addition, and perhaps most importantly, hazard modeling allows for the study of the occurrence and timing of concomitant events, such that the delicate interplay between important contextual predictors might be studied. With hazard modeling, researchers have a straightforward method of examining relationships between event occurrence and these critical time-varying descriptors.

Recommendations for the Design of Longitudinal Studies

Increase the number of waves of data collection

Strangely enough, few published studies of psychological or psychopathological "development" are truly longitudinal. Most rely upon cross-sectional or two-wave designs. Unfortunately, neither single-wave studies nor even two-wave studies provide a sufficient basis for studying development. We believe that investigators allocating limited research resources would be better served by increasing the number of waves of data collection, even at the expense of the total number of children studied.

What's wrong with cross-sectional designs? Basically, they tell us nothing about patterns of change and event occurrence. If a cross-sectional study of adolescents in a high school reveals that younger children exhibit higher levels of delinquency than their older peers, can we infer that delinquency decreases with age? Although the logical answer may be "yes," the empirical answer is a resounding "no." Even within the same school, a random sample of high school seniors will differ from a random sample of high school freshmen in potentially important ways—the two groups entered school in different years, they have experienced different significant life events, and perhaps most importantly, the sample of high school seniors will not include peers who dropped out before reaching their senior year. Observed differences in delinquency between age-separated cohorts, then, may be due to nothing more than differences in these background characteristics, not to differences in development.

Two-wave studies are only marginally better. In the case of the measurement of change, for instance, the difference between a person's observed score at Time 1 and his or her score at Time 2 can tell us whether change has occurred from beginning to end but is inadequate for studying change because it reveals nothing about the shape of each person's trajectory. Did all the change occur immediately after Time 1 or was progress steady over the entire interval? The more complex the temporal shape of the individual trajectory or the baseline hazard function, the more waves of data must be collected for the clear analytic description of that shape.

How many waves of data are enough? The advantages associated with additional waves of data collection depend, in part, upon the shape of the growth trajectory or the baseline hazard function. We must collect at least one more data point than there are unknown parameters in the individual growth model or in the baseline hazard function. In the case of the measurement of change, the adoption of a linear individual growth model, with its pair of intercept and slope parameters, requires that at least three waves of data be collected from each person under study. More complex growth models increase the data requirements—a quadratic model requires at least four waves, cubic models at least five. Similar conclusions apply in the case of hazard modeling. Such requirements imply that, to design their studies well, empirical researchers must use a combination of theory, prior research, or, better yet, pilot data, to make an educated guess about the potential shape of the growth trajectory or the hazard profile.

Whether we measure change or model event occurrence, however, these minimal requirements simply provide one degree of freedom per person for estimating model goodness-of-fit. Just because we are able to estimate a model's parameters does not imply that these parameters have been estimated well. Parameter estimation will always be improved if further waves of data are added to the design.

In the case of the measurement of change, for example, we can make the case for additional waves of data in two ways: (a) at the individual level, by examining the precision with which the change will ultimately be measured, and (b) at the group level, by considering the reliability of the change measurement. At the individual level, the precision with which we can estimate the parameters of an individual growth model improves dramatically when more waves of data are collected (see also Cook & Ware, 1983). We illustrate this in the left-hand panel of Figure 4, in which we plot the standard error with which the individual rate of change can be estimated (in units of residual standard deviation) as a function of the number of waves of data collected.[6] Notice that the relationship is strictly monotonic—as more waves of data are collected, the smaller the standard error of the estimated linear slope becomes, reflecting improved precision for the measurement of individual change. We reach the same conclusion at the group level by examining the relationship between the reliability with which change can be measured and the number of waves of data collected.[7] We display this relationship in the right-hand panel of Figure 4.

Inspection of Figure 4 also suggests that adding waves of data to an existing design gives a "bigger bang for the buck" when the original number of waves was small. This gain can be seen by examining the slopes of either of the curves in Figure 4. Notice that these slopes are steeper initially, and then decline as the number of waves of data increases. Adding an extra wave of data collection to a design that has only three waves, then, has a much greater impact on precision and reliability, proportionally speaking, than adding an extra wave to a design that has eight waves.[8]

Similar conclusions can be inferred for the estimation of the baseline hazard profile in the

[6] In Figure 4, we assume linear individual growth, independent homoscedastic normally distributed Level-1 measurement errors, and equally spaced occasions of measurement.

[7] The reliability with which change is measured is defined here as the proportion, in the population, of the observed variance in linear slope that is true variance in linear slope.

[8] Plots like Figure 4 can be used to design data collection, by permitting the investigator to decide in advance on the number of waves of data required for measurement precision or measurement reliability to reach a target level (see Singer & Willett, 1996; Willett, 1989).

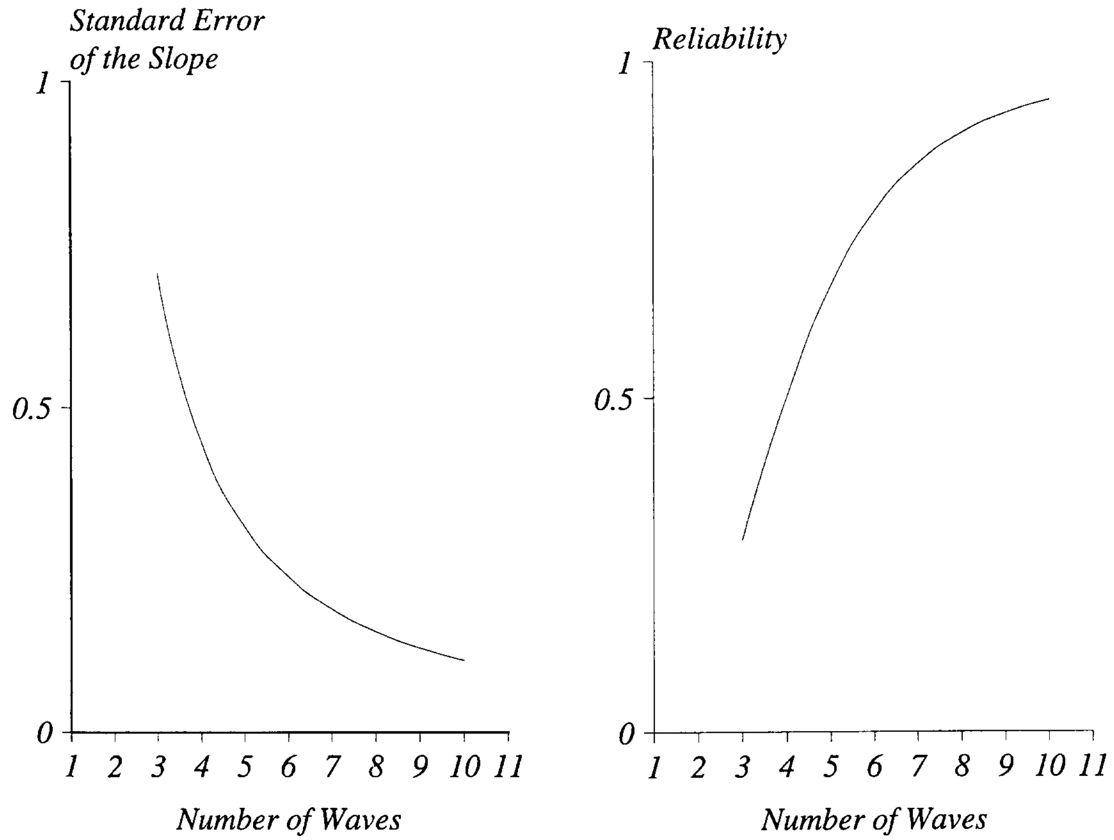

Figure 4. The benefits of increasing the number of waves of data collection. Left panel: standard error of individual rate of change (in units of residual standard deviation); right panel: reliability of rate of change. Both panels assume linear individual growth, ordinary least squares estimation of the rate of change, and equally spaced waves of longitudinal data.

case of discrete-time survival analysis. Overall, the message is clear—collect extra waves of data at all costs!

Consider accelerated longitudinal designs

Longitudinal research is not without its disadvantages. Two of the most prominent are the amount of time it takes to complete a study and the risk that its findings may be out-of-date by the time data collection (and analysis) ends. If a single cohort of 6th graders is tracked for, say, 7 years (through 12th grade), the next generation's 6th graders may behave nothing like those in the original sample when they were that young. So, too, few researchers (and funding sources) want to wait for the end of a lengthy longitudinal study before analyzing data and presenting findings.

Accelerated longitudinal designs—also known as cohort-sequential designs (Nesselroade & Baltes, 1979) or mixed longitudinal designs (Berger, 1986)—shorten the length of time needed to conduct longitudinal research. Although there are many different types of accelerated design, they all share one characteristic: rather than follow a single age-homogeneous cohort for the entire age period of interest, select two or more distinct age cohorts and track each for a shorter period of time. In the most common accelerated designs, data collection begins in a single base year. In the Adolescent Pathways Project (APP), for example, Seidman (1991) tracked two cohorts of students annually for 3 years, from 1989 to 1991, with initial data collection for each cohort beginning in those grades immediately preceding a potentially disruptive event of interest—the transition from one type of school (middle school, junior high, or high school) to the next. The younger cohort was comprised of 863 5th and 6th graders; the older cohort was comprised of 470 8th and 9th graders. By the third wave of data collection, members of the younger cohort were in 8th and 9th grade (the same grades as the members of the older cohort during the first wave of data collection) and members of the older cohort were in 10th and 11th grade. Within 3 years, Seidman had three waves of longitudinal data on students covering seven distinct grades from the 5th through the 11th.

Accelerated longitudinal designs have another advantage as well—they can help unravel the inherent confounding known as the "Age, Period, and Cohort" problem (Mason & Fienberg, 1985; Schaie, 1965). A student's place in time is marked by (a) his or her birth year ("cohort"), (b) his or her age (or grade in school), and (c) the chronological year (or "period") being described (2000, 2001, etc.). Although developmentalists emphasize the effects of age, outcomes may also be a function of the child's year of birth (the cohort effect) and the actual year being described (the period effect). Flynn (1987), for example, identified potentially profound cohort effects when he examined data from more than a dozen countries over 10- to 20-year periods and found that within less than a generation, average scores on IQ tests rose between 5 and 25 points.

The analytic problem is that all three dimensions of time are intimately linked—knowledge of any two defines the third. Data on 10-year-olds in the year 2000, for example, describe children born in 1990. This dependence makes it difficult to determine whether observed differences across individuals should be attributed to age (as is commonly done) or whether cohort and period effects also play a role. Cross-sectional studies confound the effects of age with the effects of cohort (although *age* is commonly assumed to be the overriding factor) and they preclude examination of period effects because chronological time (the year of data collection) is held constant. Traditional longitudinal studies confound the effects of age and period (although *age* is once again usually given precedence) and they preclude the examination cohort effects (because cohort is held constant by sampling). Accelerated longitudinal designs, in contrast, can provide insight into age, period, and cohort effects. By comparing parameter estimates from growth trajectories for the two cohorts in the APP, for example, Seidman could determine whether eighth graders in the younger group differ from eighth graders in the older group. Although he could not ascribe differences unequivocally to the effects of cohort (because the eighth grade data for the two cohorts were collected in different periods [years]), lack of a difference would be reasonably interpreted by most researchers as a sign of no cohort effect.

To unravel the Age, Period, and Cohort problem further, the common accelerated design can be modified in one of two ways—through the re-initiation of data collection in multiple base years (see Singer & Willett, 1996) or through a lengthening of the period of overlap between cohorts. The APP could be amplified into a multiple-base-year accelerated design by fielding a second 3-year data collection plan 1 (or 2) years after the initial round. The additional data would allow the researcher to add explicit variables representing the effects of period and cohort into the growth models and hazard models presented in Equations 2 and 3 (e.g., Raudenbush & Chan, 1992; Singer & Willett, 1988; Singer, 1993).

Alternatively, the length of the overlap between the two or more cohorts in a single-base year design can be expanded. Most accelerated designs employ a single overlapping age (or grade) set to be at the edge of both cohorts. Setting the overlap at the edge of the cohorts maximizes the length of the overall developmental trajectory, while still providing the minimal amount of overlap necessary (one wave) for piecing together distinct individual growth models and hazard functions (Anderson, 1995). But this practice has a cost. First, it limits the *precision* with which differences in the trajectories can be measured, providing the least powerful test of cohort differences possible in an accelerated design. Second, it limits the researcher to investigating only differences in level across the two cohorts, not differences in shape or slope. Lengthening the period of overlap—even a modest increase from 1 to 2 years—can reap major rewards. Had Seidman set the APP older cohort to begin with seventh and eighth graders (instead of eighth and ninth graders), for example, the overall developmental record would have been diminished modestly (from seven to six grades), but it might have been better able both to reveal cohort or period effects and facilitate tests of complex hypotheses about the shape of the growth trajectory (or hazard function).

Despite these advantages, we do not advise researchers to use accelerated designs routinely; rather, we suggest that they consider them under certain circumstances. First, these designs are most suitable when limited resources preclude the fielding of a long-term data collection effort and when interest focuses on short-term developmental issues, not long-term developmental pathways (Farrington, 1991). The piecing together of segmented growth models and hazard functions can never replace the information contained in truly longitudinal studies conducted over extended periods of time. Second, the geographic and social mobility of the communities under study must also be scrutinized—accelerated designs are most appropriate in stable environments with little migration. In- or out-migration can cause a researcher to label erroneously differences across samples as cohort effects when they are more likely attributable to preexisting, contextual differences between the groups, such as differences in socioeconomic status or cognitive ability, for example, that have nothing to do with the year that the sample members were born.

Recommendations for Measurement in Longitudinal Studies

Collect equatable data prospectively

All variables can be classified as either time invariant or time varying. In longitudinal studies of development and psychopathology, outcome variables are time varying by definition, but predictors, in contrast, may be either time varying or time invariant. Whenever time-varying variables are measured, their values must be equatable across all occasions of measurement (Goldstein, 1979), and we suggest that such data be collected prospectively and not retrospectively.

Seemingly minor differences across occasions—even those invoked to improve data quality—will undermine equatability. Changing item wording, response category labels, or the setting in which instruments are administered can render responses nonequatable. In a longitudinal study, at a minimum, item stems and response categories must remain the same over time. Although administering an identical instrument repeatedly can produce panel conditioning, empirical studies suggest that conditioning effects are small (see, e.g., Kasprzyk, Duncan, Kalton, & Singh, 1989) and their consequences pale when compared with those of measurement modification (Light, Singer, & Willett, 1990). The time for instrument modification is during pilot work, not data collection.

We also strongly recommend prospective data collection. Even simple information collected by retrospection—on the occurrence and spacing of events—can be unreliable, imprecise, and unequatable. Although important one-time events—such as age at menarche—may be remembered indefinitely, and highly salient and stressful events—such as a psychiatric hospitalization—may be remembered for several years, habitual events—such as daily activities—are forgotten almost immediately (Bradburn, 1983). Psychological states appear more prone to recall errors than do social experiences (Lin, Ensel, & Lai, 1997), but even simple questions about social states have been shown to be unreliable (Henry, Moffitt, Caspi, Langley, & Silva, 1994). The longer the period of retrospection, the greater the error—respondents forget events entirely (memory failure), remember events as having occurred more recently (telescoping), and drop fractions and report even numbers or numbers ending in 0 and 5 (rounding).

Data should be collected retrospectively only when this method of collection does not compromise their measurement. Administrative records can be invaluable in this regard as they can be used to reconstruct retrospective event histories of quality equal to those gathered prospectively. If retrospective data must be gathered directly from individuals, questionnaires must be constructed carefully. Standardized checklists are now believed to be inadequate (Raphael, Cloitre, & Dohrenwend, 1991), while life-history calendars (Freedman, Thornton, Camburn, Alwin, & Young–DeMarco, 1988; Lin et al., 1997), handheld computers (Shiffman et al., 1997), and diaries (Silberstein & Scott, 1991) have been growing in popularity. The most successful retrospective data collection strategies link questions about when an event occurred to contextual questions about where and why it happened (Bradburn, Rips, & Shevell, 1987); use narrative formats that allow the respondent, not the interviewer, to structure the course of the interview (Means, Swan, Jobe, & Esposito, 1991); and use memory aids, whenever possible, to improve recall.

Never standardize

Psychologists have a penchant for standardization. When reporting regression results, they often present standardized regression coefficients in addition to, or instead of, raw regression coefficients. When analyzing longitudinal data on the same variable over time, they often standardize the measures to mean zero and a standard deviation of one before analysis.

We understand the desire for standardization. Few psychological variables have well-accepted interpretable metrics. In comparison to economics, for example, where variables are measured on commonly understood scales (e.g., dollars, percentages), psychologists often work with variables measured in arbitrary metrics. Few experienced professionals have an intuitive understanding of what a score of, say, 15 means on even a frequently used psychological instrument, let alone one developed solely for a particular study.

Two other well-cited justifications for standardization depend upon arguments that are fundamentally flawed. One line of reasoning is that standardization helps identify the "relative importance" of predictors in a regression model (for instance, see Everitt, 1996; Marasciuolo & Serlin, 1988). The argument is that standardization eliminates the difficulties inherent in comparing regression coefficients when predictors have been measured on different scales, allowing the predictor with the largest standardized coefficient to be declared the "most important." Unfortunately, identifying the most important predictor in a statistical model is not that easy (Healy, 1990) and standardization does little to help the researcher in this regard (Bring, 1994). The other line of reasoning suggests that standardization facilitates the comparison of findings across different samples, allowing assessment of whether different studies of the same phenomenon detect effects of the same magnitude. Yet, as we show below, standardization does just the opposite, rendering it impossible to compare results across studies (Greenland, Schlesselman, & Criqui, 1986).[9]

To understand the difficulties with standardization, let us review how standardized regression coefficients are computed. Because the argument can be understood using regression models of cross-sectional data, and because the problems identified simply escalate when longitudinal data are involved, we begin with the simpler framework. Consider a regression model linking the level of delinquent behavior for individual j (DELBEH_j) to two predictors: familial rule-setting (RULES_j) and history of maltreatment ($\mathrm{MALTREAT}_j$, a dummy variable coded as 0 or 1):

$$\mathrm{DELBEH}_j = \beta_0 + \beta_1\mathrm{RULES}_j + \beta_2\mathrm{MALTREAT}_j + \varepsilon_j, \quad (5)$$

where β_1 is the population difference in delinquent behavior per unit difference in RULES,

[9] We hasten to note that applied researchers are not solely responsible for their mistaken use of standardized coefficients. We believe that the writers of statistical software (and documentation for software) encourage standardization through the misleading labeling of output. Some software packages (e.g., SPSS) use the label beta to refer to standardized regression coefficients creating the misimpression that these quantities estimate population regression parameters, given that statisticians usually write the latter using β's. In reality, the population regression parameters labeled β and the standardized regression coefficients labeled beta have little to do with each other.

controlling for maltreatment status, β_2 is the population difference in delinquent behavior between maltreated and comparison children, controlling for level of familial rule setting, and ε is a residual.

Standardized regression coefficients for this model can be obtained in one of two ways. Under the first method, each variable in the regression model is first standardized by converting to a *Z* score (by subtracting the variable's sample mean and dividing by its sample standard deviation)

$$x_j^* = \frac{(x_j - \bar{x})}{s_x},$$

and then the standardized outcome is regressed on the standardized predictor(s). Equivalently, standardized coefficients can be obtained directly by multiplying raw regression coefficients by the ratio of the sample standard deviation of the predictor to the sample standard deviation of the outcome:

$$\beta^* = \hat{\beta}\frac{s_x}{s_y}. \qquad (6)$$

In either case, the interpretation is identical—the coefficient now indicates the standardized difference in the outcome per standard deviation difference in the focal predictor, controlling for all other predictors in the model.

As the interpretation seems straightforward and the calculations seem innocuous, why do we argue that standardization is problematic? First, despite its intuitive appeal, standardization does not render the metrics of the predictors (here RULES and MALTREAT) comparable. All that has happened is that the predictors have been transformed to a mean of 0 and a standard deviation of 1. What does it mean for two substantively distinct variables to possess a common standard deviation? Whenever one or more of the predictors is a dichotomy (as in this example), such interpretation is near impossible. In this situation, standardization actually destroys the intuitively appealing interpretation of β_2 in Equation 5 replacing it with a convoluted interpretation involving the standard deviation of a variable that can only take on the values 0 and 1. Even if the two (or more predictors) are continuous, standardization does not render unit differences in the variables comparable. Is a one standard deviation difference in rule setting the same as a one standard deviation difference in a variable like maternal education? The answer to this question depends upon the sample homogeneity with respect to these variables, which in turn depends, at least in part, on researchers' decisions about target populations and sampling strategies. Yet standardization effectively eliminates information about homogeneity from consideration, creating the false illusion that coefficients can be directly compared.

In any statistical model, the only coefficients that can be compared directly are those for which the predictors have been measured in identical units. If one predictor describes the number of hours that a child spends with family members and another predictor describes the number of hours that a child spends with friends, a researcher can compare these predictors' *raw* coefficients to evaluate the effect of an extra hour of family time versus an extra hour of peer time. Even in this situation, however, the standardized coefficients present little new information and tell us nothing about which variable is more important in predicting delinquent behavior.

Standardized regression coefficients are not only unhelpful, they can be misleading. If the standard deviation of either the outcome or any of the predictors differs across samples, samples with identical population parameters (the true values of the regression coefficients in Equation 5) can yield strikingly different standardized regression coefficients creating the erroneous impression that results differ across studies. So, too, samples with distinctly different population parameters can yield identical standardized regression coefficients creating the erroneous impression that the results are *similar* across studies. Moreover, the discrepancy between the raw and standardized regression coefficients can be in either direction (larger or smaller), providing no rule-of-thumb for evaluating the size of the underlying effect. The bottom line is that differences across samples in the standard deviations of either the predictors or the outcome can lead to mistakes about similarities or differences of effects.

Lest one think that this type of sample-to-sample difference is a theoretical contrivance unlikely to happen in practice, several simple "thought experiments" suggest the opposite. Even when sampling from the same target population, for example, different random samples will have different standard deviations, with the differences being potentially more pronounced when sample sizes are small (as when studying rare populations). When sampling from different target populations, the probability of different standard deviations escalates, increasing the probability that standardized regression coefficients will differ when true regression coefficients are the same and that they will be similar when true regression coefficients differ. This discrepancy is especially likely when comparing samples recruited using different strategies—say, one from the schools and another from hospitals—or when one researcher studies a normative sample and the other studies a clinical one.

As a corollary to this point, differences in study design can also affect the magnitude of standardized regression coefficients. In the model presented in Equation 5, the standardized regression coefficient for MALTREAT will differ depending upon its standard deviation, which in turn is directly related to the proportion of maltreated children under study. All else being equal, as the percentage of maltreated children departs from 50% (in either direction), the standard deviation will decrease, producing a decrease in the standardized regression coefficient. As a result, studies that compare two groups using balanced designs will yield standardized coefficients that are larger than identical studies using unbalanced designs, even if the true mean difference between the groups is identical. Similarly, two studies can yield identical standardized coefficients even when the true mean difference between groups is anything but identical.

The problems associated with standardization escalate in longitudinal studies. Not only do the issues outlined above continue to apply, but if the researcher decides to standardize within waves of data (as is common), two additional problems emerge. First, standardizing the outcome within-wave places unnecessary and unusual constraints on its variation. If the collection of individual growth curves fans out over time (as is common), standardizing the outcome within-wave essentially increases the amount of outcome variation manifest during early time periods and diminishes the amount of outcome variation manifest during later ones. The resulting standardized growth trajectories bear little resemblance to the raw trajectories and may even mislead the researcher into thinking that growth is nonlinear when it is actually linear, or vice versa (Willett, 1985). Second, because all longitudinal studies suffer some attrition, standardization of predictors within waves inevitably relies on means and standard deviations that are estimated in a decreasing pool of subjects (as is especially the case when studying the occurrence of events in atypical, high-risk populations such as psychiatric inpatients). If attrition is nonrandom (and it usually is), then the successive samples used in the estimation of the predictor means and standard deviations will be nonequivalent, and the standardized values of the predictors will be noncomparable from wave to wave. Thus, for example, if the level of antisocial behavior is studied over time in a group of hospitalized boys, many of whom drop out of treatment, move on to a different type of facility, or refuse to comply with future data gathering efforts, standardized predictor values cannot be compared from one wave to the next due to the nonrandom nature of the loss of subjects from each successive sample. Given that developmental psychopathologists are often concerned with studying the process of adaptive and maladaptive behavior among high risk samples, such a caution about standardization within longitudinal studies is especially pertinent.

Recommendations for the Analysis of Longitudinal Data

Consider alternative specifications for the effect of time

Time is the fundamental predictor in both individual growth modeling and hazard modeling strategies. So, models specified under either approach must include at least one predictor that represents the effect of time. Although researchers typically devote their intellectual energy to modeling the effects of *substantive* predictors (e.g., parental divorce, child maltreatment), we believe that there are benefits to paying more attention to the modeling the effect of this *structural* predictor, time. Most investigators assume that the underlying growth model is linear in time, and fail to investigate the possibility that an alternative temporal structure might be more realistic and more substantively appealing. So, too, most researchers assume that the baseline hazard function ($\beta_0(t)$) is best represented as a step-function, failing to investigate the possibility that a simpler, smoother function might suffice.

Why is it important to specify the effects of time appropriately? From a substantive perspective, the answer is simple—the specification of "the effect of time" describes the shape of the underlying developmental trajectory. Is growth linear or nonlinear? Does the hazard function peak, and if so, when? Are the developmental trajectories smooth and continuous, or are there jumps corresponding to time-linked events in the child's life? Contemplation of alternative specifications for the effect of time creates an array of modeling options that, if used appropriately, can lead to more accurate summaries of complex development and facilitate the testing of interesting hypotheses about the effects of substantive variables. This linkage is perhaps easiest to appreciate in the context of growth modeling, where the individual growth parameters defined in a level-1 model (as in Equation 1) become the conceptual outcomes in a level-2 model (as in Equation 2). Specifying the effect of time in a sensible way at level-1 ensures that the individual growth parameters have meaningful substantive interpretations—at the simplest level, perhaps as an initial status and a rate of change. These parameters then become the outcomes at level 2, permitting investigation of links between initial status and rate of change on the one hand and contextual characteristics of the individual on the other. Changing the specification of time at level 1 alters the interpretation of the individual growth parameters and the nature of the questions that can be addressed at level 2. Finally, if an alternative specification is indeed a more appropriate representation of reality, the model will better fit the data and the statistical power of associated hypothesis testing will improve.

What alternative specifications of the effect of time might we consider? We begin our discussion of alternative specifications in the context of individual growth modeling and then describe how these ideas extend to the case of hazard modeling. Although the array of possible specifications is endless, we limit ourselves to suggesting four that, taken together, provide a sense of the opportunities available. For simplicity, we present models for analyzing data from a hypothetical five-wave study that follows students annually from 6th to 10th grade. For the first three waves of data collection (6th–8th grade) the students are in junior high; for the remaining two waves (9th–10th grades), they are in high school. This simplification allows us to treat the variable time synonymously with grade.

Figure 5 presents four possible alternative specifications for the effect of time. Model A (top left panel) presents a variant of the classic linear individual growth model, in which child j's score in grade i is expressed as a weighted linear combination of an intercept (π_{0j}), a slope (π_{1j}) and an error term unique to that student on that occasion (ε_{ij}). In the same way that the intercept in Equation 1 was defined by centering on age 11, notice that the temporal predictor here is GRADE-8. As in Equation 1, the subtraction of eight from every value of GRADE creates our interpretation of the intercept (π_{0j}) as child j's true status on Y in eighth grade. The interpretation of the linear slope π_{1j} has its usual interpretation as the annual rate of change.

Manipulating the interpretation of the intercept (by recentering the predictor representing time) is the easiest and most common modification of an individual growth model. Although the researcher can recenter in many ways, we chose to illustrate the general idea by subtracting eight here because eighth grade is (a) the mid-point of data collection (allowing the intercept to be interpreted as the "average value of Y" during the study period); and (b) a substantively meaningful point in time (the last year of junior high). A subsequent level-2 model exploring variation in these level-1 parameters would now identify predictors of eighth grade status (and growth), a desirable property for researchers more interested in status in eighth grade than in sixth. Willett (1997) extends this idea further by showing how these models can be expressed using combinations of final status and growth, and even final status and initial status.

The remaining reparameterizations for the effect of time presented in Figure 5 involve three or more growth parameters, and therefore have steeper data requirements (as noted earlier). Model B (top right panel) is a piecewise linear growth model, another extension of Model A. Because the 5-year data collection period tracks students from junior high through high school, this model adds a *shift* parameter, π_{2j}, which indicates the differential in *Y* that kicks in when child *j* graduates from junior high. The resultant growth trajectory is comprised of two linear segments (hence the name piecewise) that are parallel (guaranteed through the use of the single slope for the variable GRADE-8) but that differ in level before and after graduation (represented by the dichotomous predictor, JHSGRAD, coded 0 before graduation and 1 after).

Piecewise linear growth models are useful when a researcher expects a sudden discontinuity in the growth trajectory at a known point in time. In the most common applications, the shift coincides with a substantive transition—changing grades, graduating from school, or seasonal fluctuations. In an analysis of five waves of data collected on students between the spring of first grade and the spring of third grade, for example, Bryk and Raudenbush (1988) used this type of model with a variable labeled SUMMER DROP, which registered the decline in achievement test scores that occurred each fall after summer vacation. Piecewise linear models can also be used when analyzing longitudinal data collected on participants before and after an experimental manipulation, with the shift parameter registering the implementation of the innovation.[10]

The piecewise linear growth model has a major, and sometimes unrealistic, constraint: the segments before and after the shift are assumed to be parallel. Model C (bottom left panel of Figure 5) relaxes this constraint, allowing the two segments to differ not only in level but also in slope. Interpretation of π_{0j}, π_{1j}, and π_{2j} remain the same. The additional parameter π_{3j} indicates the difference in growth rates between the two periods of time. Positive values of π_{3j} indicate a steeper slope in high school; negative values a shallower one.

Allowing for the possibility of differential slopes before and after a transition or intervention represents a substantial leap forward in temporal parameterization. The size and sign of the slope differential parameter (π_{3j}) provides a direct glimpse of the effects of context on development. A non-zero value indicates that growth rates differ during different phases of children's lives. Growth in prosocial activity might be rapid in preadolescence and slower during the teen years, while growth in risk-taking behavior might be gradual in preadolescence and rapid thereafter. Moreover, the companion level-2 models allow researchers to ask whether the differential in growth rates varies systematically across children as a function of individual, familial, or environmental contextual characteristics. Is the escalation in risk-taking behavior similar across children, or is it especially pronounced for those living in single-parent households or inner city neighborhoods? Although use of this model requires five or more

[10] A further extension of the piecewise linear growth model includes a transition coded to correspond to events in the individual's life. In a study of body image over time, for example, a researcher might include a predictor labeled MENARCHE in the level-1 model. Because events like these are not only time-varying, the periodicity with which they change differs across people, making interpretation of parameters somewhat more difficult. Nevertheless, although we do not explore such models here, we hasten to add that they are interesting and important extensions of these ideas.

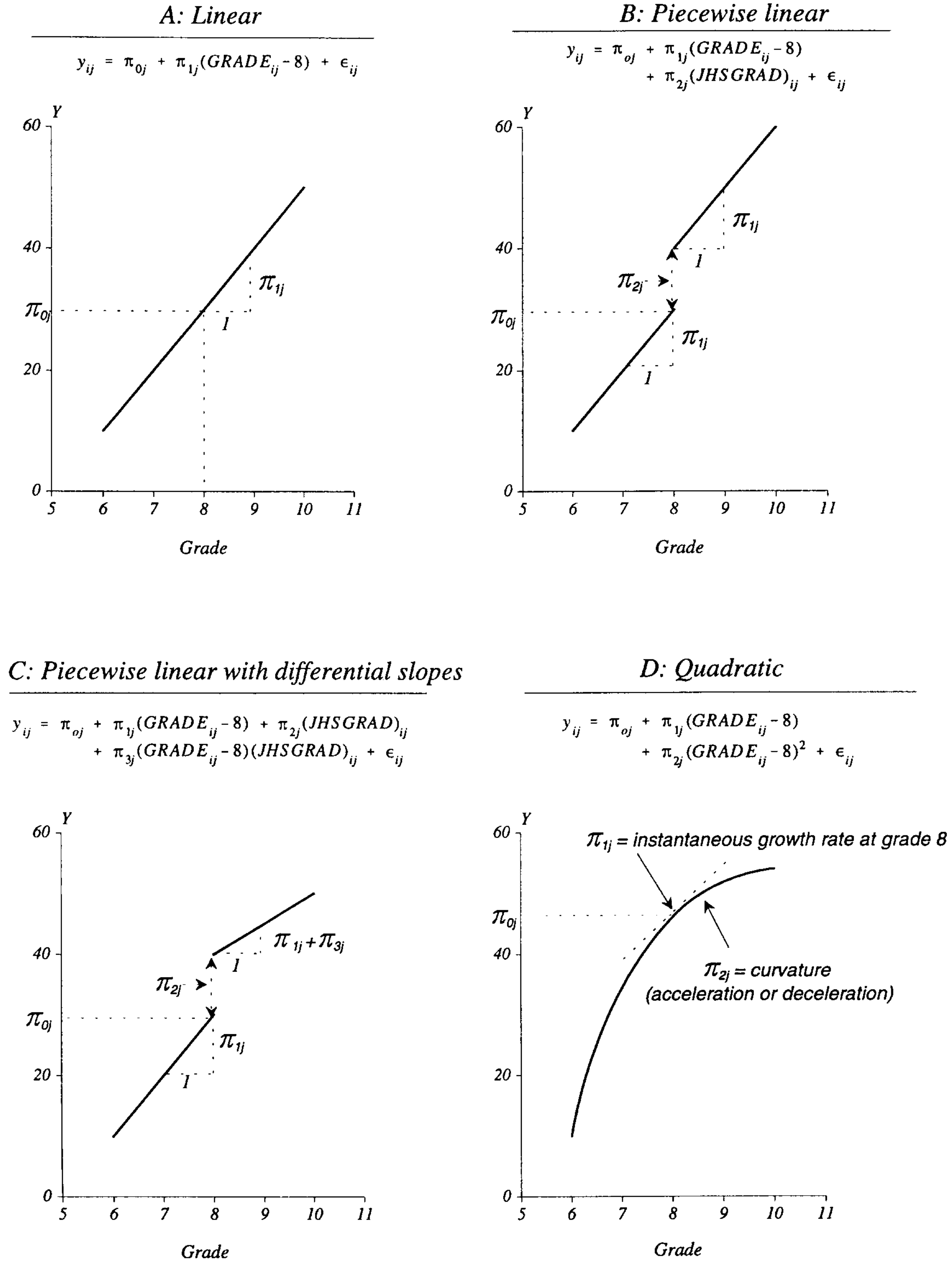

Figure 5. Alternative parameterizations for the effect of time in individual growth models. The four panels present an array of specifications for time in the level-1 (within-person) individual growth model: (A) linear, (B) piecewise linear, (C) piecewise linear with differential slopes, and (D) quadratic.

waves of data, we believe that its generality facilitates exploration of some fascinating substantive hypotheses about development.

The quadratic individual growth model (Model D, bottom right panel of Figure 5) also allows growth rates to differ across the life span, but unlike the other three models it assumes that changes in the slope are smooth—that is, that individual growth is nonlinear. The addition of the squared predictor GRADE-8^2 to the linear model in Model A permits the growth rate to differ smoothly and systematically as a function of age. Special care is needed when interpreting parameters in nonlinear models. Although π_{0j} still indicates the true value of Y for child j in eighth grade, π_{1j}, the coefficient on (GRADE-8), is now the instantaneous rate of true growth in grade 8—that is, the slope of a tangent to the curve at eighth grade. The sign and the size of the curvature parameter (π_{2j}) indicates the manner and degree to which the growth curve departs from a straight line. If $\pi_{2j} = 0$, there is no curvature—the model is linear. If π_{2j} is negative, the curve decelerates over time (as shown in the graph)—the greater the absolute value of π_{2j}, the greater the deceleration. If π_{2j} is positive, the resultant curve would be flipped over (top to bottom), and growth would be accelerating over time, with larger values indicating more acceleration.

Quadratic growth models share the same substantive advantages as the piecewise linear growth model with differential slopes. They, too, permit examination of relationships between differentials in growth rates and contextual characteristics of participants, their families, and their communities. Do the growth trajectories for abused children mirror those of control children, or do abused children accelerate at a slower rate? But unlike Model C, which constrains growth to be linear over the short term with differential slopes in different phases of life, the quadratic model assumes that growth rates differ smoothly and continuously as a function of age. Piecewise linear models may suffice in short-term studies with little growth, but when studying a rapidly changing outcome, or when studying people over a long period of time, quadratic (and even higher order polynomial) models may provide a better fit. Nonlinear specifications have two further advantages as well. First, because they contain fewer parameters, they have less costly minimum data requirements. The quadratic model, for example, can be fit with just four waves of data. Second, they can easily be extended to more complex trajectories through the use of higher order polynomials. Willett (1997) provides a detailed discussion of these and related ideas.

So far, we have discussed alternative specifications for the effect of time only in the context of individual growth modeling. Although not immediately obvious, all the parameterizations we have discussed can also be used in discrete-time hazard models. When we introduced the hazard model in Equation 3, we purposefully did not indicate any specific parameterization for time. Instead, we indicated the baseline hazard profile as $\beta_0(t)$, a completely general representation. We did so because, historically, researchers using the two different methods have differed in their approach to parameterization. Those using growth modeling typically begin (as we have) with a linear formulation, adapting to more complex representations only as necessary. Those fitting discrete-time hazard models, in contrast, typically begin with a completely general specification for the time predictor—a step function represented by a series of dummy variables, one per time period—and then explore smoother and less complex parameterizations.

We illustrate these ideas using 11 years of longitudinal data collected by Widom (1989) in her study of the sequelae of childhood abuse and neglect. Keiley and Martin (1998) reanalyzed these data using discrete-time hazard modeling to predict whether and, if so, when subjects were first arrested for a juvenile offense. Although less than one-quarter of the children were arrested before age 18, a pronounced developmental pattern was associated with age at first arrest. It is this pattern that we explore here. Figure 6 presents two alternative baseline hazard models depicting the risk of first arrest as a function of child age. The fainter segmented line represents the baseline hazard function estimated using a

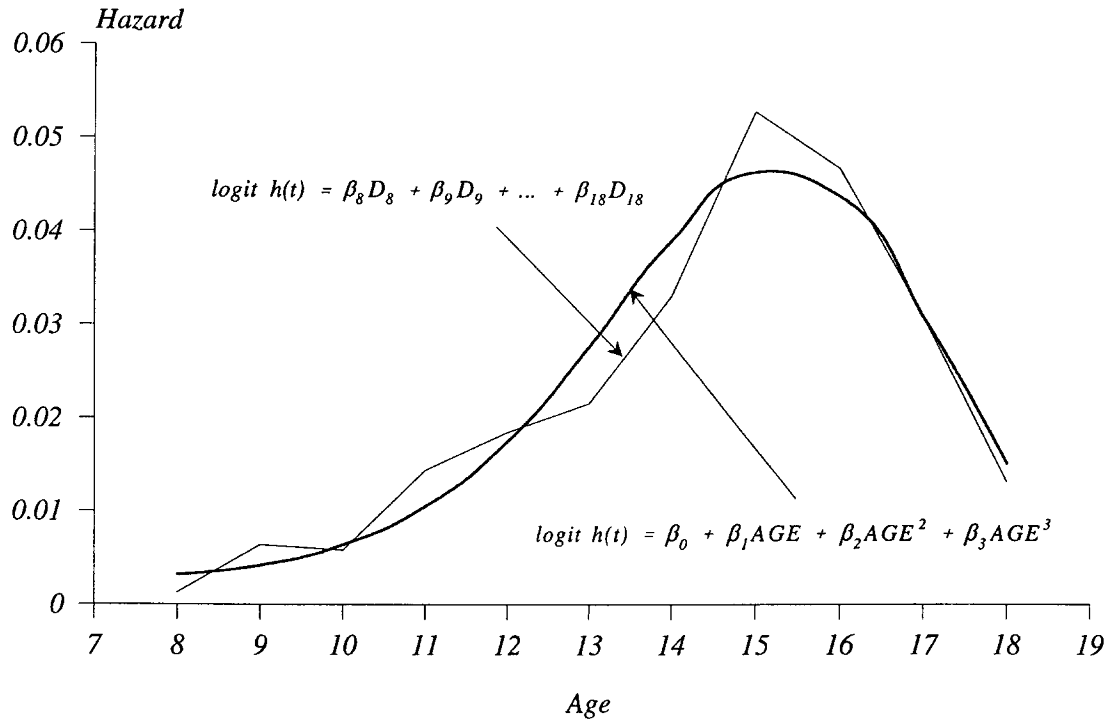

Figure 6. Alternative parameterizations for the effect of time in hazard models. Fitted hazard functions describing the age at first juvenile arrest using smooth (darker curve) and totally general (lighter joined lines) specifications for the effect of time.

completely general representation for $\beta_0(t)$. As shown in the figure, this model expresses logit hazard as a linear function of 11 dummy variables, D_8–D_{18}, at one per year. Each coefficient β then represents the value of logit hazard in that time period—β_8 represents logit hazard for 8-year-olds, β_9 for 9-year-olds, and so on. We plot the fitted hazard function by transforming each of these coefficients using the standard formula for reexpressing logits: hazard = $1/\{1 + e^{-\beta}\}$. Notice that hazard is minuscule in the preteen years, rises during early adolescence, and peaks at age 15 years. After that, the risk of first arrest among those who have not yet been arrested declines. The smooth parameterization—the darker curve—replaces the general specification with a cubic spline (also known as a second-order polynomial). In place of 11 dummy variables representing time, three continuous variables—AGE, AGE^2, and AGE^3—now represent the effect of time. Notice how well the smooth curve approximates the jagged one. Keiley and Martin (1998) show that the smooth function (which requires only four parameters for specification) is preferable to the completely general representation (which requires eleven) because the goodness-of-fit statistic associated with the former closely approximates that associated with the latter, using far fewer parameters. We present a more detailed discussion of these types of model comparisons in Willett and Singer (1993) and Singer and Willett (1993).

In practice, how can a researcher select an appropriate specification for the effect of time? We believe that at least four issues should be considered. First, different specifications have different data requirements. The more parameters involved, the more waves of data needed. Although researchers fitting hazard models typically have sufficient data to explore the completely general formulation presented in Figure 6, those fitting growth models often work within tighter constraints. (This is one reason why we recommend that researchers extend the length of their longitudinal investigations beyond the three waves currently considered the norm.) Second, researchers should ask whether theory may suggest a particular functional form. Does theory suggest that growth is smooth, or does it suggest that the outcome changes in fits and starts? Do the gaps correspond to particular events in the child's life, or do they occur seemingly at random? Third, does previous research suggest a particular functional form? In studies of human lifetimes, for example, where hazard models are used routinely, the shape of the baseline hazard function is so well established that researchers typically adopt particular parametric forms (e.g., Lee, 1996). Fourth, what functional form do the data suggest? Examination of empirical growth trajectories and sample hazard functions computed separately by values of each predictor can often suggest reasonable places to begin.

Always test for interactions, particularly between substantive predictors and time

Social scientists often display a "main effects" bias. When fitting statistical models, they explore the main effect of each predictor, perhaps alone and after controlling statistically for the effects of other predictors. But why should predictors operate only as main effects? Many developmental theories identify important contextual predictors whose effects should differ systematically across people, across places, or across the life span. For example, the effect of restrictive parenting on the cognitive functioning of children from high-risk families (i.e., families within high-crime areas) has been shown to be positive, perhaps offering a protective buffer of a sort, while such parenting in low-risk families has been proven to be counterproductive to children's cognitive development (Baldwin, Baldwin, & Cole, 1990).

Whenever the effect of one predictor differs by levels of another predictor, we say that the two predictors interact.[11] Interactions are powerful tools for exploring subtle (and not so subtle) differences in how individuals react under seemingly similar circumstances and thus for studying the differential effects of context on developmental outcomes. For example, in line with the concept of multifinality within the field of developmental psychopathology (Cicchetti, 1990), why do some contexts not always lead to the same psychopathological outcome, such that some children are resilient when faced with a parental divorce, while others descend into a cycle of psychopathology? Why do some individuals respond positively to prevention programs, while others remain resistant? Addressing these types of research questions requires the inclusion of interaction terms in statistical models.

Researchers who do explore interactions typically focus on interactions among substantive predictors. In their study of the development of schizophrenia, for example, Walker and colleagues (1996) explore interactions between stress and a variety of physiological and psychosocial predictors. But researchers investigating growth and event occurrence can explore a more fascinating type of interaction: the interaction with time. When a predictor interacts with time, its impact on the outcome is different in different time periods. By exploring interactions with time, a researcher can determine whether a predictor's effect (e.g., parental attachment) remains the same across the life span, or whether its effect fluctuates as individuals age. In a study of depression in adolescents, for example, the effect of family factors on depression may decline as children mature while the effect of peer factors may increase.

Interactions with time can perhaps be best understood via an example that compares the effects of two predictors—one that does not interact with time and one that does. To focus, we return to the discrete-time hazard models specified for the first onset of depression data earlier. Recall that when we introduced the time-varying predictor representing parental divorce (PARDIV(t)), we examined the relationship between it and the risk of depression (Equation 4; Figure 3). PARDIV(t) is a time-varying predictor—its value goes from 0 to 1 if, and when, parents divorce. Now, we ask if its effect on hazard is really constant over time (as we have stipulated so far). If the effect is time invariant, then the impact of pa-

[11] Although statisticians prefer the term interaction, many psychologists refer to these situations by saying that the action of one predictor *moderates* the effect of another.

rental divorce on the risk of onset is the same regardless of whether the divorce takes place during childhood, adolescence, or adulthood. If the effect of parental divorce differs over time, in contrast, divorce may have a larger effect on the risk of depression among children, who are still living at home, say, than among adults, who have already moved out of the house.

But now we appear to have modeling dilemma. The discrete-time hazard models posited in Equations 3 and 4 do not permit a predictor's effect to differ with time. In these models, *proportional-odds models,* the hazard profiles have a special property: in every time-period (t) under consideration, the effect of the predictor on logit-hazard is exactly the same. In Equation 3, for example, the vertical shift in the logit-hazard profile for women is always β_1 and, consequently, the hypothesized logit-hazard profiles for women and men have identical shapes, since their profiles are simply shifted versions of each other. Generally, in proportional-odds models, the entire family of logit-hazard profiles represented by all possible values of the predictors share a common shape and are mutually parallel, differing only in their relative elevations. If the logit-hazard profiles are parallel and have the same shape, the corresponding raw hazard profiles are (approximate) magnifications and diminutions of each other—they are proportional.[12] Because the models presented so far include predictors with only time-constant effects, the fitted hazard functions displayed appear to have the required proportionality.

But what if the effects of some predictors are not time-constant? What if some hazard profiles corresponding to different values of the predictor are not proportional to each other? Many predictors will not only displace the logit-hazard profile, they will alter its shape. If the effect of a predictor varies over time, we must specify a nonproportional model that allows the shapes of the logit-hazard profiles to differ. To include such an effect in our hazard models, we simply include the cross product of that predictor and time as an additional predictor.

To illustrate the types of conclusions that can be gleaned from testing whether a predictor interacts with time, Figure 7 presents the results of fitting two discrete-time hazard models to the depression data but introducing a new predictor of one aspect of family context—number of siblings (NSIBS).[13] Because NSIBS is a continuous variable (whose values vary from 0 to 26), we present fitted hazard profiles for two prototypical individuals: those who were only children (no siblings) and those who came from larger families (six siblings). The figure presents fitted hazard profiles from two distinct models: a main-effects model (top panel) and an interaction-with-time model (bottom panel). The main effects model suggests that siblings protect against depression: for both men and women, the greater the number of siblings, the lower the risk of onset. The four fitted hazard profiles appear proportional because the main effects model constrains the effect of NSIBS to be the same in each time period.

But a more accurate and complex story emerges from the interaction-with-time model displayed in the bottom panel, in which the effect of NSIBS is allowed to vary over time. Comparing the fitted hazard functions from the interaction with time model with those from the main effects model illustrates the untenability of the proportionality assumption, due to the statistically significant interaction between NSIBS and time. The hazard functions in the bottom panel are clearly not proportional. In childhood, when individuals are still living at home, family size does have a protective effect: boys and girls from larger families are at lower risk of having an initial depressive episode. Over time, however, the protective effect of family size diminishes and

[12] For pedagogic reasons, we have taken mathematical liberties here. In discrete-time models, the hazard probability is usually small (say, less than .15 or .20). When discrete-time hazard is about this magnitude, or less, the approximation tends to hold quite well (see Singer & Willett, 1993, for further discussion).

[13] Because of data limitations, the values of this predictor are assumed to be constant during an individual's lifetime. If we knew when the respondent's siblings were born, we could have coded this predictor as time varying.

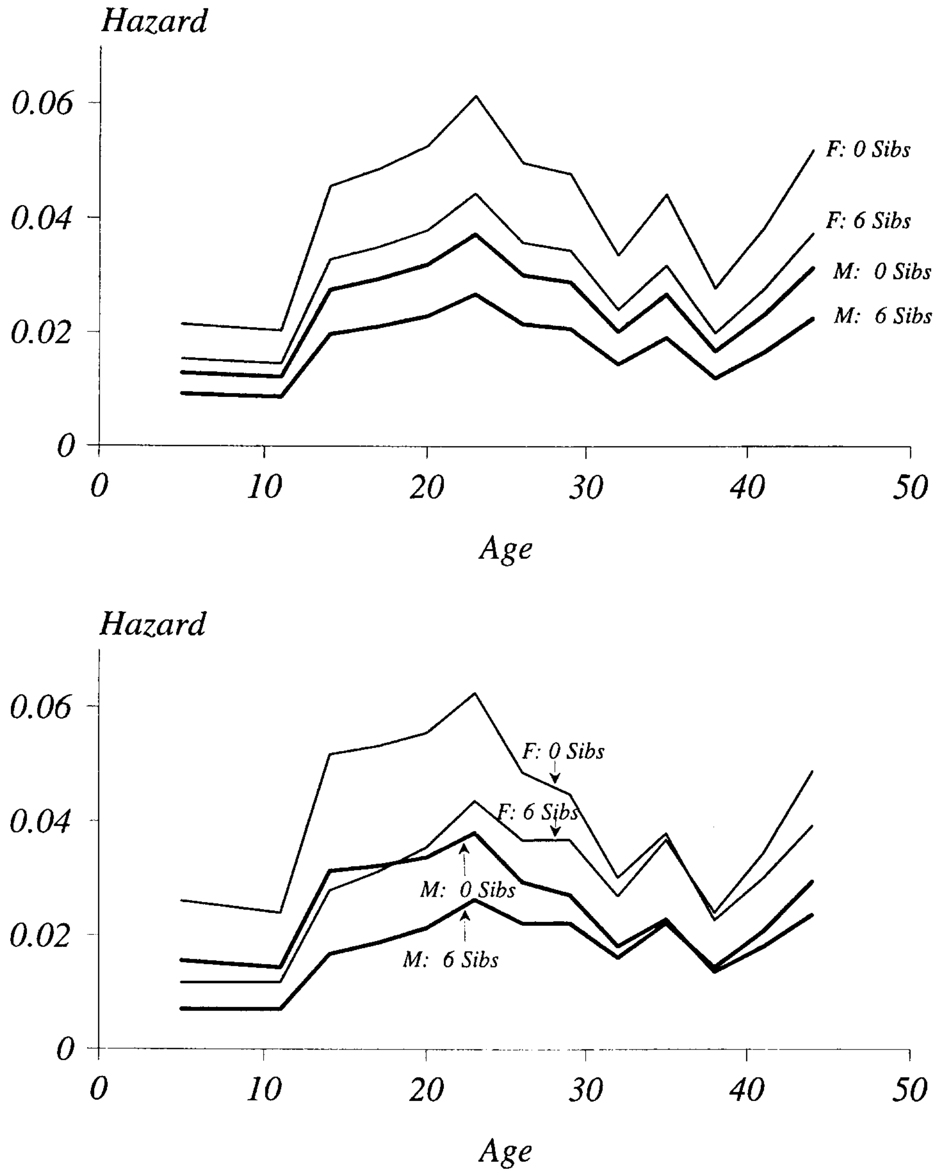

Figure 7. Including an interaction with time in hazard models. Fitted hazard functions describing the age at first onset of depression, by gender and the individual's number of siblings, from two discrete-time hazard models. The top panel is a main effects model in which the effect of number of siblings is *constant* over time; the bottom panel is an interaction-with-time model in which the effect of number of siblings *varies* over time.

by the time an individual reaches his or her early thirties, the effect is virtually nonexistent. Instead of having a constant vertical separation in logit-hazard space, the relative differences between the hazard functions differ, being larger in childhood and trivial in adulthood. Simply put, the effect of family size on the risk of depression interacts with time.

The ability to include, and test the importance of, interactions with time as predictors in growth models and hazard models represents a major analytic opportunity for researchers investigating the effects of context on development. When studying the behavior of individuals over very long periods of time, it is logical to ask whether the effects of important contextual predictors vary as people pass through different life stages. Although the effects of some predictors will remain unchanged with an individual throughout his or her lifetime, the effects of others may dissipate, or increase, over time.

We believe that interactions with time are everywhere and would be found more often if researchers systematically and intentionally looked for them. Present data analytic practice (and the widespread availability of prepackaged computer programs) permits an almost unthinking (and often untested) adoption of proportional hazards models (Cox regression), in which the effects of predictors are constrained to be constant over time. We have found, however, in a wide variety of substantive applications including not only our own work on employment duration (Murnane, Singer, & Willett, 1989; Singer, 1993a, 1993b) and age at entry into day care (Singer et al., in press) but also others' work on topics such as age at first suicide ideation (Bolger et al., 1989) and child mortality (Trussel & Hammerslough, 1983), that interactions with time seem to be the rule, rather than the exception. The key is to test the tenability of the assumption of a time-invariant effect. For a description of analytic methods for achieving this, we refer the reader to Singer and Willett (1993) and Willett and Singer (1993).

How do these ideas extend to the case of growth models? Although it is not immediately obvious, the traditional individual growth model specified in Equations 1 and 2 actually *assumes* an interaction with time. This can be seen most clearly by substituting the level-2 equations back into the level-1 equation to yield the single combined model:

$$\begin{aligned} Y_{ij} = \beta_{00} + \beta_{10}(\text{AGE}_{ij} - 11) + \beta_{01}(\text{FEMALE}_j) \\ + \beta_{11}(\text{FEMALE}_j)(\text{AGE}_{ij} - 11) \\ + \{u_{0j} + u_{1j}(\text{AGE}_{ij} - 11) + \varepsilon_{ij}\}. \quad (7) \end{aligned}$$

Notice that the structural part of the combined model (presented on the first line of Equation 7) includes three predictors—a main effect of AGE, a main effect of FEMALE, and a cross product of FEMALE and AGE. This cross product term represents the interaction between FEMALE and AGE. The parameter β_{11}, therefore, represents the magnitude of the difference in growth rates for boys and girls, and a test of its statistical significance indicates whether the predictor (here FEMALE) interacts with time. When we fit this model to the NLSY data, we found that the growth rates did not differ by gender; in other words, there was no statistical interaction. Were this a substantive paper, we would therefore modify the level-2 equation in Equation 2 to eliminate the effect of FEMALE on the level-1 slope parameter (π_{1j}) and refit the model to data. If prototypical growth trajectories were then plotted by gender, we would see a pair of parallel lines—one for boys and one for girls—with identical slopes, illustrating that the effect of gender did not interact with time (cf. Figure 1, right-hand panel).

Postscript

We believe that the methods of individual growth modeling and survival analysis present exciting opportunities in answering many of the developmental questions with which developmental psychopathologists and others are concerned. These methods allow for the analysis of longitudinal data and thus facilitate the process of uncovering the various pathways along which development may occur. Such a pathways approach is essential for developmental psychopathologists who seek to study the course of both maladaptive and adaptive behavior throughout the life span (Cicchetti, 1993). In addition, the methods outlined in this paper allow for the incorporation of any number of predictors of development, including important contextual predictors such as family structure, school environment, or neighborhood climate, and thus provide psychopathologists with the tools both to study the range of outcomes that may be associated with a particular set of predictors (multifinality) and to explore how similar outcomes may result from a variety of contexts (equifinality). It is our hope that by adopting the statistical methods discussed in this paper and following the guidelines we suggest, the study of developmental psychopathology and, in particular, the study of development in the context in which it occurs, will be enhanced.

References

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data.* Sage University Paper Series on Quantitative Applications in the Social Sciences (Serial No. 05-046). Beverly Hills, CA: Sage.

Anderson E. R. (1995). Accelerating and maximizing information from short-term longitudinal research. In J. Gottman (Ed.), *The analysis of change* (pp. 139–164). Mahwah, NJ: Erlbaum.

Baldwin, A. L, Baldwin, C., & Cole, R. E. (1990). Stress-resistant families and stress-resistant children. In J. Rolf, A. S. Masten, D. Cicchetti, K. H. Nuechterlein, & S. Weintraub (Eds.), *Risk and protective factors in the development of psychopathology* (pp. 257–280). Cambridge, UK: Cambridge University Press.

Belle, D., Norell, S., & Lewis, A. (1997). Becoming supervised: Children's transitions from adult-care to self-care in the afterschool hours. In I. H. Gotlib & B. Wheaton (Eds.), *Stress and adversity over the life course: Trajectories and turning points* (pp. 159–178). Cambridge, UK: Cambridge University Press.

Berger, M. P. (1986). A comparison of efficiencies of longitudinal, mixed longitudinal, and cross-sectional designs. *Journal of Educational Statistics, 11*(3), 171–181.

Bolger, N., Downey, G., Walker, E., & Steininger, P. (1989). The onset of suicide ideation in childhood and adolescence. *Journal of Youth and Adolescence, 18,* 175–189.

Bradburn, N. M. (1983). Response effects. In P. H. Rossi, J. D. Wright, & A. A. Anderson (Eds.), *Handbook of survey research* (pp. 289–328). San Diego, CA: Academic Press.

Bradburn, N. M., Rips, L. J., & Shevell, S. K. (1987). Answering autobiographical questions: The impact of memory and inference on surveys. *Science, 236,* 157–161.

Bring, J. (1994). How to standardize regression coefficients. *The American Statistician, 48*(3), 209–213.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101,* 147–158.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models in social and behavioral research: Applications and data analysis methods.* Newbury Park, CA: Sage.

Cicchetti, D. (1990). An historical perspective on the discipline of developmental psychopathology. In J. Rolf, A. Masten, D. Cicchetti, K. Nuechterlein, & S. Weintraub (Eds.), *Risk and protective factors in the development of psychopathology* (pp. 2–28). New York: Cambridge University Press.

Cicchetti, D. (1993). Developmental psychopathology: Reactions, reflections, projections. *Developmental Review, 13,* 471–502.

Cicchetti, D., & Rogosch, F. A. (1996). Equifinality and multifinality in developmental psychopathology. *Development and Psychopathology, 8,* 597–600.

Collett, D. (1991). *Modeling binary data.* London: Chapman & Hall.

Cook, N. R., & Ware, J. H. (1983). Design and analysis methods for longitudinal research. *Annual Review of Public Health, 4,* 1–23.

Dodge, K. A., Pettit, G. S., & Bates, J. E. (1994). Effects of physical and maltreatment on the development of peer relations. *Development and Psychopathology, 6,* 43–55.

Everitt, B. S. (1996) *Making sense of statistics in psychology.* New York: Oxford University Press.

Farrington, D. P. (1991). Longitudinal research strategies: Advantages, problems, and prospects. *Journal of the American Academy of Child and Adolescent Psychiatry, 30,* 369–374.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171–191.

Freedman, D., Thornton, A., Camburn, D., Alwin, D., & Young-DeMarco, L. (1988). The life-history calendar: A technique for collecting retrospective data. *Sociological Methodology, 18,* 37–68.

Friedman, H. S., Tucker, J. S., Schwartz, J. E., & Tomlinson-Keasey, C. (1995). Psychosocial and behavioral predictors of longevity: The aging and death of the "Termites." *American Psychologist, 50,* 69–78.

Goldstein, H. (1979). *The design and analysis of longitudinal studies: Their role in the measurement of change.* New York: Academic Press.

Gore, S., Aseltine, Jr., R., Colten, M. E., & Lin, B. (1997). Life after high school: Development, stress, and well-being. In I. H. Gotlib & B. Wheaton (Eds.), *Stress and adversity over the life course: Trajectories and turning points* (pp. 197–214). Cambridge, UK: Cambridge University Press.

Gotlib, I. H., & Wheaton, B. (1997). *Stress and adversity over the life course: Trajectories and turning points.* Cambridge, UK: Cambridge University Press.

Greenland, S., Schlesselman, J. J., & Criqui, M. H. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology, 123*(2), 203–208.

Hagan, J., & McCarthy, B. (1997). Intergenerational sanction sequences and trajectories of street-crime amplification. In I. H. Gotlib & B. Wheaton (Eds.), *Stress and adversity over the life course: Trajectories and turning points* (pp. 73–90). Cambridge, UK: Cambridge University Press.

Healey, M. J. R. (1990). Measuring importance. *Statistics in Medicine, 9,* 633–637.

Hedeker, D., Gibbons, R., & Flay, B. R. (1994). Random effects regression models for clustered data with an example from smoking prevention research. *Journal of Consulting and Clinical Psychology, 62,* 757–765.

Henry, B., Moffitt, T. E., Caspi, A., Langley, J., & Silva, P. A. (1994). On the "remembrance of things past": A longitudinal evaluation of the retrospective method. *Psychological Assessment, 6,* 92–101.

Kandel, D. B., & Davies, M. (1986). Adult sequelae of adolescent depressive symptoms. *Archives of General Psychiatry, 43,* 255–262.

Kasprzyk, D., Duncan, G. J., Kalton, G., & Singh, M. P. (1989). *Panel surveys.* New York: Wiley.

Keiley, M. K., & Martin, N. C. (1998). The relationship between childhood abuse and later adolescent delinquency. Manuscript submitted for publication.

Kreft, I. G. G., de Leeuw, J., & van der Leeden, R. (1994). Review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, ML3, VARCL. *The American Statistician, 48,* 324–335.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design.* Cambridge, MA: Harvard University Press.

Lin, N., Ensel, W. M., & Lai, W. G. (1997). Construction and use of the life history calendar: Reliability and validity of recall data. In I. H. Gotlib & B. Wheaton (Eds.), *Stress and adversity over the life course: Trajectories and turning points* (pp. 249–272). Cambridge, UK: Cambridge University Press.

Marasciulo, L. A., & Serlin, R. C. (1988). *Statistical methods for the social and behavioral sciences*. New York: Freeman.

Mason, W. M., & Fienberg, S. E. (1985). *Cohort analysis in social research: Beyond the identification problem.* New York: Springer–Verlag.

Means, B., Swan, G. E., Jobe, J. B., & Esposito, J. L. (1991). An alternative approach to obtaining personal history data. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 167–183). New York: Wiley.

Murnane, R. J., Singer, J. D., & Willett, J. B. (1989). The influences of salaries and "opportunity costs" on teachers' career choices: Evidence from North Carolina. *Harvard Educational Review, 59*(3), 325–346.

Nesselroade, J. R., & Baltes, P. B. (Eds.). (1979). *Longitudinal research in the study of behavior and development.* New York: Academic Press.

Nolen–Hoeksema, S. (1990). *Sex differences in depression.* Stanford, CA: Stanford University Press.

Petersen, A. C., Sarigiani, P. A., & Kennedy, R. E. (1991). Adolescent depression: Why more girls? *Journal of Youth and Adolescence, 20,* 247–272.

Raphael, K. G., Cloitre, M., & Dohrenwend, B. P. (1991). Problems of recall and misclassification with checklist methods of measuring stressful life events. *Health Psychology, 10*(1), 62–74.

Raudenbush, S. W., & Chan, W. (1992). Growth curve analysis in accelerated longitudinal designs. *Journal of Research in Crime & Delinquency, 29*(4), 387–411.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 90,* 726–748.

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika, 50*(2), 203–228.

Schaie, W. K. (1965). A general model for the study of developmental problems. *Psychological Bulletin, 64,* 92–107.

Seidman, E. (1991). Growing up the hard way: Pathways of urban adolescents. *American Journal of Community Psychology, 19,* 173–205.

Shaw, D. S., Owens, E. B., Vondra, J. I., Keenan, K., & Winslow, E. B. (1996). Early risk factors and pathways in the development of early disruptive behavior problems. *Development and Psychopathology, 8,* 679–699.

Shiffman, S., Hufford, M., Hickcox, M., Paty, J. A., Gnys, M., & Kassel, J. D. (1997). Remember that?: A comparison of real-time versus retrospective recall of smoking lapses. *Journal of Consulting and Clinical Psychology, 65,* 292–300.

Silberstein, A. R., & Scott, S. (1991). Expenditure diary surveys and their associated errors. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 303–326). New York: Wiley.

Singer, J. D. (1993a). Are special educators' career paths special?: Results of a 13-year longitudinal study. *Exceptional Children, 59,* 262–279.

Singer, J. (1993b). Once is not enough: Former special educators who return to teaching. *Exceptional Children, 60*(1), 58–72.

Singer, J. D. (in press). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics.*

Singer, J. D., Fuller, B., Keiley, J., & Wolf, A. (in press). Early child care selection: Variation by geographic location, maternal characteristics, and family structure. *Developmental Psychology.*

Singer, J. D., & Willett, J. B. (1988). Uncovering involuntary layoffs in teacher survival data: The year of leaving dangerously. *Educational Evaluation and Policy Analysis, 10,* 212–224.

Singer, J. D., & Willett, J. B. (1991). Modeling the days of our lives: Using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. *Psychological Bulletin, 110,* 268–290.

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics, 18,* 155–195.

Singer, J. D., & Willet, J. B. (1996). Methodological issues in the design of longitudinal research: Principles and recommendations for a quantitative study of teachers' careers. *Educational Evaluation and Policy Analysis, 18,* 265–283.

Sorenson, S. B., Rutter, C. M., & Aneshensel, C. S. (1991). Depression in the community: An investigation into age of onset. *Journal of Consulting and Clinical Psychology, 59*(4), 541–546.

Tremblay, R. E., Masse, L. C., Vitaro, F., & Dobkin, P. L. (1995). The impact of friends' deviant behavior on early onset of delinquency: Longitudinal data from 6 to 13 years of age. *Development and Psychopathology, 7,* 649–667.

Trussel, J., & Hammerslough, C. (1983). A hazards-model analysis of the covariates of infant and child mortality in Sri Lanka. *Demography, 20,* 1–26.

Walker, E. F., Neumann, C. C., Baum, K, Davis, D. M., DiForio, D., & Bergman, A. (1996). The developmental pathways to schizophrenia: Potential moderating effects on stress. *Development and Psychopathology, 8,* 647–665.

Wheaton, B., Roszell, P., & Hall, K. (1997). The impact of twenty childhood and adult traumatic stressors on the risk of psychiatric disorder. In I. H. Gotlib & B. Wheaton (Eds.), *Stress and adversity over the life course: Trajectories and turning points* (pp. 50–72). Cambridge, UK: Cambridge University Press.

Widom, C. S. (1989). Child abuse, neglect, and adult behavior: Research design and findings on criminality, violence, and child abuse. *American Journal of Orthopsychiatry, 59*(3), 355–367.

Willett, J. B. (1985). *Investigating systematic individual differences in academic growth.* Unpublished doctoral dissertation, Stanford University, Palo Alto, CA.

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Rothkopf (Ed.), *Review of research in education 1988–89* (pp. 345–422). Washington, DC: American Educational Research Association.

Willett, J. B. (1994). Measurement of change. In T. Husen & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd. ed.). Oxford, UK: Elsevier Science Press.

Willett, J. B. (1997). Measuring change: What individual growth modeling buys you. In E. Amsel & K. A. Renninger (Eds.), *Change and development* (pp. 213–243). Mahwah, NJ: Erlbaum.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116*(2), 363–381.

Willett, J. B., & Singer, J. D. (1993). Investigating onset, cessation, relapse, and recovery: Why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *Journal of Consulting and Clinical Psychology, 61,* 952–965.

Willett, J. B., & Singer, J. D. (1995). It's deja-vu all over again: Using multiple-spell discrete-time survival analysis. *Journal of Educational and Behavioral Statistics, 20*(1), 41–67.

Yamaguchi, K. (1991). *Event history analysis.* Newbury Park, CA: Sage.