Annu. Rev. Psychol. 2001. 52:501–25

COMPARING PERSONAL TRAJECTORIES AND DRAWING CAUSAL INFERENCES FROM LONGITUDINAL DATA

Stephen W. Raudenbush
School of Education and Institute for Social Research, University of Michigan, 610 East University Avenue, Ann Arbor, Michigan 48109; e-mail: rauden@umich.edu

■ **Abstract** This review considers statistical analysis of data from studies that obtain repeated measures on each of many participants. Such studies aim to describe the average change in populations and to illuminate individual differences in trajectories of change. A person-specific model for the trajectory of each participant is viewed as the foundation of any analysis having these aims. A second, between-person model describes how persons very in their trajectories. This two-stage modeling framework is common to a variety of popular analytic approaches variously labeled hierarchical models, multilevel models, latent growth models, and random coefficient models. Selected published examples reveal how the approach can be flexibly adapted to represent development in domains as diverse as vocabulary growth in early childhood, academic learning, and antisocial propensity during adolescence. The review then considers the problem of drawing causal inferences from repeated measures data.

CONTENTS

INTRODUCTION

Longitudinal studies serve many purposes in psychology. They trace the course of normal growth, identify risk factors for mental illness, and assess the effects

of interventions. They identify the timing of onset of new abilities, the duration of disorders, and desistance from self-destructive behavior. What these purposes share in common is a focus on individual change and thus a need for repeated assessment of each participant. The time series for a person creates a basis for sketching that person's history. Whereas a person's history can capture many domains of change—for example, changes in cognitive skill, emotional self-regulation, mood, and social behavior—a trajectory describes a person's development in one well-defined domain. In this article, I view longitudinal analysis as beginning with the development of a model for individual trajectory. Only if the trajectory is clearly defined does it become possible quantitatively to summarize evidence across trajectories. The logical foundation for all longitudinal analysis is thus a statistical model defining parameters of change for the trajectory of a single participant. The task of comparing people then becomes the task of comparing the parameters of these personal trajectories. A model is thus needed for the population distribution of the parameters of personal change.

Within this framework, I review current methodological developments that have dramatically expanded the analytic choices available to psychologists. The explosive growth of new methodology enables astonishing flexibility for studying change and comparing trajectories of change in the presence of missing data and varied timing and spacing of time points. This growth in new methods can also cause seemingly terminal confusion. Part of the confusion emanates from a proliferation of jargon often associated with new software packages. So part of my task is to clarify the new vocabulary and connect it to key logical issues in the analysis.

A brief article cannot comprehensively review a topic as vast as methods of longitudinal analysis. Therefore, the first section of this article, while outlining a variety of conceptions of "trajectory," selects a subset of these for more intensive discussion. By choosing to focus on just a part of the terrain, it becomes possible to portray quite vividly what is meant quantitatively by a trajectory and how such trajectories can be compared.

Researchers have often criticized cross-sectional research designs as inadequate for identifying causal effects and have advocated longitudinal data collection as a solution. The second half of this article therefore considers the assumptions required to justify causal inferences in longitudinal studies.

Trajectories of Change

Longitudinal studies use quite varied conceptions of a trajectory of change. Let us briefly sketch three of these before moving to a fourth conception that provides the basis for the remainder of the paper.

Models for Time as an Outcome Many important research questions concern the timing of an event, for example, the age of onset of menarche, time until recovery from depression, and the time until a student receives a degree. The probability that the event of interest will occur at time t, given that it has not already occurred,

is called the "hazard" of the event. The hazard changes over time. A participant's changing hazard is then that person's trajectory. The trajectory may depend on time-invariant covariates, including gender, social background, and personality characteristics, and also on time-varying covariates, that is, events and experiences that unfold over time. The key methodological challenge arises from the fact that the study will typically end before some members of the sample experience the event. Of those not experiencing the event, many or all will ultimately experience it after the termination of the study. For these people, the time of the event is "censored," that is, unobservable. Survival analysis involves an important class of longitudinal models that are tailored to deal with duration time as an outcome when outcomes for some participants are censored. This important class of models, reviewed, for example, by Cox & Oakes (1984), Kalbfleisch & Prentice (1980), and Singer & Willett (1991) will not be considered further in this article.

Transitions Between States It is often of interest to develop theories that predict transitions between states, for example, whether a recovering alcoholic will be actively drinking at any time t (yes or no) given that this participant was (or was not) drinking at time $t - 1$. In this kind of study, a person's trajectory is the sequence of such transitions. The probability of a given state (e.g. drinking at time t) given the prior state (e.g. not drinking at time $t - 1$) is called a transition probability. A model for a person's trajectory begins with a theory about what determines a sequence of transition probabilities for a given person. Covariates describing personal background and experience may determine these probabilities and thus strongly influence a person's trajectory. The current article will not consider these models in further detail. We refer the interested reader to Diggle et al (1994, Chapter 10), Collins (2001), and Collins & Wugalter (1992).

Sequences of Age-Appropriate Markers Developmental psychologists are interested in whether early impulsivity predicts conduct disorder in elementary school and juvenile crime later (Moffitt 1993). Status attainment researchers consider how parental social status predicts a person's educational attainment and how these, together, predict that person's adult occupation and income (Duncan et al 1972, Sewell & Hauser 1975). In these cases, the focus is on associations among variables, each of which marks a person's status at a given stage of life. A trajectory in such studies is the sequence of values on the variables of interest. Common methods of analysis include structural equation models or path models, as introduced by Blau & Duncan (1967), Blalock (1969), and Duncan (1966) and refined considerably since (cf Joreskog & Sorbom 1979). Although longitudinal data are essential to studying these important relationships, the analytic methods are not different from multivariate methods commonly used in cross-sectional research, so we shall not consider them further.

Smooth Curves to Describe Trajectories of Growth or Change Perhaps the most common analytic goal in current longitudinal psychological research is to

describe and compare growth curves or other smooth functions that describe a person's changing status. For this purpose, analysts have recommended approaches with the labels "covariance component models," "hierarchical linear models," "latent curve analysis," "mixed models," "multilevel models," "random-coefficient models," and "random-effect models." Much of the remainder of this article aims to clarify the logic of these models while giving some meaning to these labels. Researchers interested in these approaches generally collect repeated measures of a given outcome variable for each participant with the aim of describing how people grow or change over some fairly broad interval of time.

Other Studies of Repeated Measures This review is confined to studies performed over some significant segment of the life course. Such a focus omits other important types of repeated-measure studies. Studies of repeated measures over the short term, for example, can assess sequences of social interaction in detail, model reaction times in a series of cognitive tasks, or predict biophysical changes in response to a series of stimuli. We refer the interested reader to Gottman (1995, Part II), for reviews of methods for such data.

Describing and Comparing Growth and Change Curves

We begin with a simple and easily understandable example and then extend this logic to more complex cases.

Defining the Trajectory for Person i Suppose that each person's status with respect to cognitive outcome, *Y*, grows with age at a constant rate and that these rates vary randomly over a population of persons. More specifically, we have an outcome Y_{ti} for person i at time t, with T_i time points ($t = 1, \ldots, T_i$) observed for person i. We can formulate a simple linear model such as the following for individual change:

$$Y_{ti} = \pi_{0i} + \pi_{1i} a_{ti} + e_{ti}, \quad e_{ti} \sim N(0, \sigma^2) \tag{1}$$

where a_{ti} is the age of person i at time t. Here person i's trajectory is the straight line $\pi_{0i} + \pi_{1i} a_{it}$. The parameters of individual i's trajectory are two: π_{0i}, the status of that person at $a_{ti} = 0$, and π_{1i}, that person's linear rate of change per unit increase in age. The discrepancy between the actual score Y_{ti} and the trajectory is e_{ti}, assumed for simplicity here to be distributed normally with a mean of zero and constant variance, σ^2.

The simple model of Equation 1 has found varied application in psychology. Francis et al (1991) defined π_{1i} as the rate of recovery of patient i's cognitive functioning after suffering a head injury. For Raudenbush and Chan (1993), π_{1i} was adolescent i's rate of increase in antisocial thinking. For Seltzer et al 1994, π_{1i} was the rate of growth of i's reading comprehension. Having defined the parameters of personal change, the next stage of the model expresses a theory about individual

differences: differences, for example, in the rate of recovery from head injury, in the rate of increase of antisocial thinking, or the rate of growth in reading.

Describing the Population Distribution of Personal Trajectories Next we consider how the personal trajectories of change are distributed in the population. A very simple model states that person-specific parameters (π_{0i}, π_{1i}) vary around their grand means (β_{00}, β_{01}), with variance (τ_{00}, τ_{11}) and covariance (τ_{01}), according to a bivariate normal distribution:

$$\begin{aligned} \pi_{0i} &= \beta_{00} + u_{0i} \\ \pi_{1i} &= \beta_{10} + u_{1i} \\ &\textit{with} \quad \begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix} \right]. \end{aligned} \tag{2}$$

Unpacking the Terminology

The model described by Equations 1 and 2 is a "hierarchical" model because it specifies a model for Y_{ti} given first-level parameters (π_{0i}, π_{1i}), while these parameters, in turn, depend on second-level parameters (the βs and τs). Levels could be added; the second-level parameters could depend on third-level parameters, and so on. Thus, it is the hierarchical dependence among the parameters that is decisive in making the model "hierarchical," not necessarily the hierarchical structure of the data, although the two often go together.

This idea is central to the current article. All of the models considered in this paper are hierarchical. The essence of the decision framework I propose is to ensure that specification of the model at each level fits the research question, design, and data.

The model described by Equations 1 and 2 is a hierarchical linear model (Bryk & Raudenbush 1992) because, at the first stage, Y is a linear function of π_{0i}, π_{1i}, whereas at the second stage, each π is a linear function of β_{00}, β_{10}. Note that all polynomial models at the first stage are linear. The model is "multilevel" (Goldstein 1995) because it describes data that vary at two levels: within persons and between persons. The model is a "random-coefficients" model (Longford 1993) because the level-1 model defines coefficients π_{0i}, π_{1i}, that vary randomly over participants at level 2. It is a "latent-curve" model (Meredith & Tisak 1990) because the trajectory or curve $\pi_{0i} + \pi_{1i}a_{it}$ is unobservable, depending, that is, on unobserved latent variables π_{0i} and π_{1i}. It may also be a "latent-growth" model, although the use of the term "growth" implies a monotonic increasing trajectory which may or may not be the case.

The model is also a "mixed" model (Diggle et al 1994). To see this, substitute Equation 1 into Equation 2, yielding the combined model

$$Y_{ti} = \beta_{00} + \beta_{01}a_{ti} + \varepsilon_{ti} \tag{3}$$

where

$$\varepsilon_{ti} = u_{0i} + u_{1i}a_{ti} + e_{ti}. \quad (4)$$

Thus, the model has fixed effects (β_{00}, β_{01}) and random effects (u_{0i}, u_{1i}) as well as the elemental residual e_{ti}. Models that incorporate both fixed and random effects have historically been labeled "mixed models." This is a random-effects model (Laird & Ware 1982) because individual differences are characterized by random effects u_{0i}, u_{1i}. It is a covariance components model because it allows the random effects to covary.

The model is also a "structural-equation" model (SEM) (Joreskog & Sorbom 1979). The first-stage model can be viewed as a measurement model with observed variable Y, latent variables π, and factor loadings 1 (for π_{0i}) and a_{ti}, (for π_{1i}). The second-level model is an exceedingly simple structural equation model for the πs.

A Framework for Analytic Choices

The essence of the framework I propose is as follows.

1. In light of the developmental theory at hand, the design, and the data, choose a model for individual development over time. The model has a structural part that specifies a trajectory in terms of person-specific parameters (the πs). It may be a linear trajectory (including polynomials) as in Equation 1 or nonlinear. The model also includes a probabilistic part that describes the random behavior of Y around the person-specific trajectory. The probability model may be normal, as in Equation 1, or it may not.
2. In light of key research questions and data, define a model for the distribution of trajectories in the population. It will also have a structural part, which describes the expected trajectory given the measured characteristics of persons, and a probabilistic part, which describes the random behavior of the trajectories in the population.
3. From this point, we continue to higher levels as necessary. For example, the model might have a three-level structure with level 1 describing individual change, level 2 describing interindividual variation within clusters such as schools (Bryk & Raudenbush 1992, Chapter 8), and level 3 describing variation between clusters.
4. Assess alternative estimation methods, algorithms, and (hopefully) available software for making inferences about all unknowns.

Unfortunately, constraints in estimation theory, algorithms, and software often distort model choices. In particular, the convenience of linear models and normal distribution theory at each level, contrasted with the relative difficulty of constructing efficient computational algorithms for nonlinear models and non-normal assumptions, has encouraged an unhealthy reliance on linearity and normality.

Below we consider two classes of models. In the first, both the data Y (and the πs) are normal with a linear structure. In the second class, either Y (given the

πs) or the πs or both are non-normal (and possibly nonlinear). The first class of models is most familiar to psychology, and, although it has many useful applications, it is often chosen more out of convenience than conviction. The second class is enormously broad and offers vast untapped potential for the modeling of developmental phenomena.

Normal Data and Normal Trajectories

We consider the examples in which assumptions of normality and linearity are benign in this section, while urging analysts to consider the broader range of models described in the subsequent section.

Vocabulary Growth During the Second Year of Life Following the tradition of Bryk & Weisberg (1977) and Rogosa et al (1982), this article asserts that the appropriate foundation for longitudinal analysis is a model for the individual trajectory. To illustrate, we review the study by Huttenlocher et al (1991) of expressive vocabulary during the second year of life. These researchers visited the home of each child in their study every 2 months, each time counting the number of unique words uttered by the child. Figure 1 displays the scatter plot of vocabulary (vertical

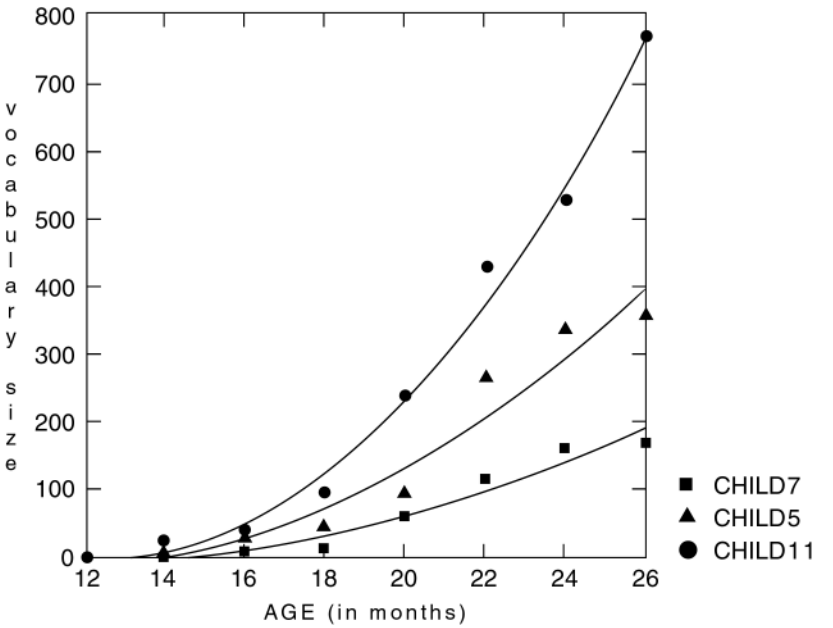


Figure 1 A sample of individual vocabulary growth trajectories. The symbols □, ◯, and △ represent the actual observations. The smooth curves result from fitting a separate quadratic polynomial to each child's vocabulary data.

axis) as a function of age in months (horizontal axis) for three randomly selected children. An accelerating curve nicely summarizes the growth pattern for each child. Such a quadratic curve can be described by the level-1 model for the growth of individual *i*:

$$Y_{ti} = \pi_{0i} + \pi_{1i} a_{it} + \pi_{2i} a_{it}^2 + e_{ti}, \tag{5}$$

where age a_{ti} is expressed as a deviation from some meaningful constant. For example, if $a_{ti} = \text{age} - 18$, months for child *i* at time *t*, π_{0i} is the size of child *i*'s vocabulary at 18 months, π_{1i} is the rate of growth of child *i* at that same age, and π_{2i} captures that child's rate of acceleration. In Huttenlocher's study, π_{2i} turned out to be the key growth parameter because acceleration was the single feature of growth that strongly differentiated the growth curves. By using all of the data to strengthen inference about correlates of acceleration, the researchers were able to discover reliable associations between key predictors (e.g. gender and maternal speech) and vocabulary growth.

Growth of Reading and Math Achievement During the Elementary Years A key feature of Huttenlocher's study was the construction of an outcome variable (vocabulary) measured on the same scale at every age of interest. Without such an invariant outcome metric, discussion of quantitative change or growth is meaningless. Standardized tests often fail to provide such a metric; different forms of a test are constructed with age-appropriate items for different age groups, and no effort is made to equate the forms. However, by calibrating the items across alternate forms, it is possible to construct a common measure for studies of cognitive growth. The Sustaining Effects Study (Carter 1984) used such a measure to characterize growth in early-elementary reading and math.

Two features of individual growth were apparent. First, academic-year growth was essentially linear. Second, growth rates during the summer were much smaller than growth rates during the academic year. Figure 2 plots the average trajectory for reading and math during the elementary years. Notice that assessments twice annually (fall and spring) enable one to distinguish academic-year and summer growth. On average, there is no growth during the summer in math, but the academic-year growth rate is substantially positive. For reading, the summer rate is modestly positive, with faster academic-year growth. To capture key features of growth, a sensible model for the individual trajectory for either math or reading is

$$Y_{ti} + \pi_{0i} + \pi_{1i}(grade_{ti} - 1) + \pi_{2i}(fall_{ti}) + e_{ti} \tag{6}$$

where $\text{grade}_{ti} - 1$ takes a value of 0 for first grade, 1 for second grade, etc, and "*fall*" is an indicator for the fall time point ($fall = 1$ if the first occurred in the fall, 0 if spring). Under this specification, π_{1i} is the calendar-year growth rate, and $\pi_{1i} + \pi_{2i}$ is the summer growth rate. Bryk & Raudenbush (1992, Chapter 8) were able to discern the following features of early elementary growth by using such an approach: (*a*) much greater academic-year than summer growth in math

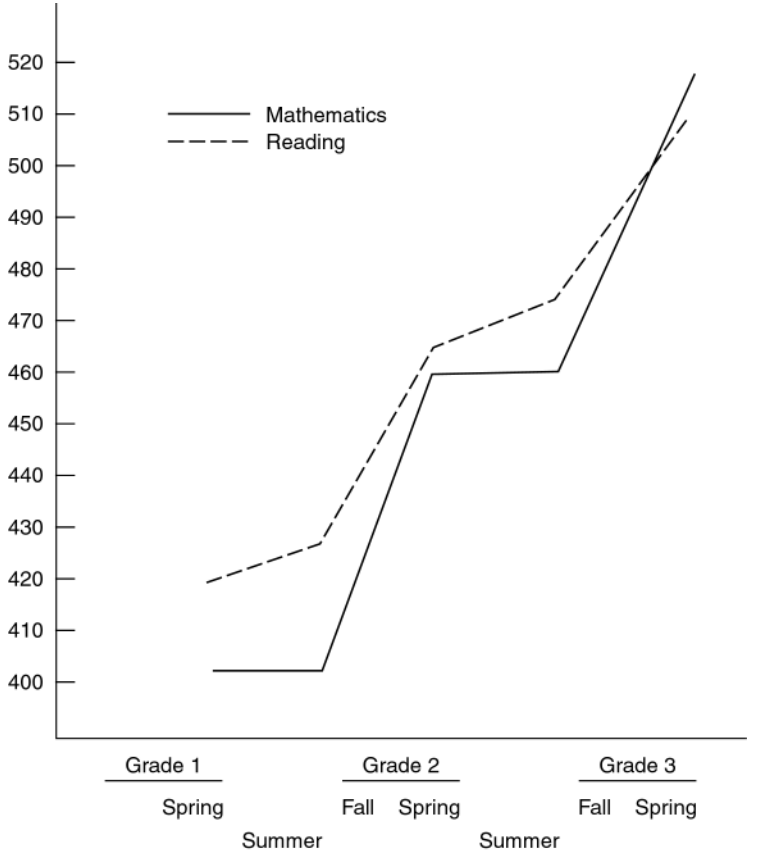


Figure 2 Expected growth functions in math and reading based on three-level model analysis of data from the sustaining effects study.

and reading on average; (*b*) modest summer growth in reading with no summer growth on average in math; (*c*) subject matter differences in school effects such that most of the variation in academic-year rates in math was between schools while most of the variation in academic-year growth in reading lay between children within schools; (*d*) a stronger dependence of growth rates on social background for reading than for math.

Summary The two examples illustrate the centrality of the model of the individual trajectory for studying longitudinal data on growth or change. For the first example, representing the accelerating nature of vocabulary growth during the second year of life was essential not only in describing individual growth and mean growth but also in discovering correlates of growth. In the second example, distinguishing between the summer growth rate and the academic-year growth rate for a

given child was essential in illuminating school effects and clarifying differences between the growth processes for reading and math.

Expanding the Class of Level-1 Models

So far we have considered only linear models to characterize a person's trajectory in the level-1 model. The level-1 model also assumes a normally distributed random error e_{ti}. The generalized linear model of McCullagh & Nelder (1989) provides a far more general class of level-1 models.

Consider Horney et al's (1995) longitudinal study of high-rate offenders. Their interest focused on how changing life circumstances such as getting a job or moving in with a spouse are related to the propensity to commit crime. They therefore conceived each participant's propensity as having a trajectory over time that could be deflected by such changes in life circumstances. However, the data, collected by means of a life history calendar over 36 months, involved binary Y_{ti}; that is, $Y_{ti} = 1$ if person i committed a crime during time t, $t = 1, \ldots, 36$. The binary character of the data strongly suggests that a Bernoulli probability model

$$\begin{aligned} E(Y_{ti} \mid \mu_{ti}) &= Prob(Y_{ti} = 1 \mid \mu_{ti}) = \mu_{ti} \\ Var(Y_{ti} \mid \mu_{ti}) &= \mu_{ti}(1 - \mu_{ti}). \end{aligned} \tag{7}$$

Note that, unlike the normal model, the conditional variance at level 1 is intrinsically heteroscedastic, depending on μ_{ti} and therefore varying over i at time t.

A linear structural model for μ_{ti} would make little sense in this case. A linear model would be inconsistent with the bounds ($0 < \mu_{ti} < 1$) because a linear model could easily generate predicted values less than zero or greater than one. Thus, any effect sizes associated with such a model would be suspect. A more natural model and that used by Horney et al is the logit-linear structural model

$$\eta_{ti} = \log\left(\frac{\mu_{ti}}{1 - \mu_{ti}}\right) = \pi_{0i} + \sum_{p=1}^{P} \pi_{pti} a_{pti}. \tag{8}$$

Here a_{pti}, $p = 1, \ldots, P$, are time-varying predictors, measured aspects of life circumstances; and μ_{pi} can be thought of as a deflection in the propensity to commit crime associated with a change in a_{pti}. At level 2, the πs can be viewed as having the same kind of model as before (Equation 2), that is, a linear structural model and a multivariate normal-probability model.

In sum, the two-level model of Horney et al (1995) is similar conceptually to Equations 1 and 2, conceiving each participant to have a trajectory of development characterized by certain parameters π that, in turn, vary over a population of persons. However, the nature of the data calls for a nonlinear structural model at level 1 and a non-normal sampling model.

In the language of the generalized linear model (McCullagh & Nelder 1989), the logit transformation η_{ti} of the mean μ_{ti} is called the link function for the binomial mean. Other standard link functions are the log-link for count data and the identity

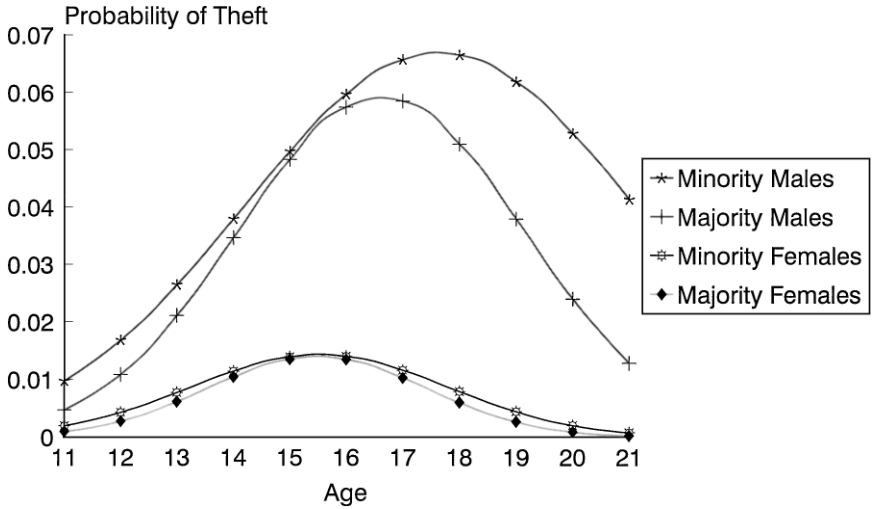


Figure 3 Probability of theft during adolescence.

link for normal data. Typically, the link function is set equal to a linear model as in Equation 8; that is, η_{ti} is a linear function of the πs. Diggle et al (1994) review applications of hierarchical generalized linear models.

However, it will not always be the case that a linear model for the link function captures the interesting features of development. Consider now the "age-crime" curve, described by Gottfredson & Hirschi (1990) as one of the "brute facts" of criminology. Researchers have found that, for many societies and many subgroups, the probability of committing a serious crime tends to be very small during preadolescence. However, this probability tends to increase at an accelerating rate early in adolescence, typically reaching a peak around ages 16–17 and then diminishing rapidly during late adolescence and early adulthood. This curve thus takes on a bell shape as a function of age, as displayed in Figure 3, based on an analysis of data from all seven cohorts of the National Youth Survey (Raudenbush 2001).

These curves describe the fitted values based on a two-level hierarchical model. At level 1, the outcome Y_{ti} takes on a value of 1 if participant i commits serious theft during time interval t, and the value is 0 if not. We thus adopt a Bernoulli probability model as in Horney et al (1995) and, provisionally, a logit linear structural model in

$$\eta_{ti} = \pi_{0i} + \pi_{1i} a_{ti} + \pi_{2i} a_{ti}^2, \tag{9}$$

which the log odds of committing serious theft are a quadratic function of age: where a_{ti} is the age of person i at time t minus 16 (the median age of the sample during the 5 years of data collection). Ages range from 11 to 21 across the seven cohorts during the 5 years of the study. At level 2, each of the three πs depends on gender (an indicator for female status), family income, and ethnicity (an indicator

for minority status, with 1 = African American or Hispanic American; 0 = other):[1]

$$\pi_{pi} = \beta_{p0} + \beta_{p1}(\textit{female})_i + \beta_{p2}(\textit{minority})_i + u_{pi}, \tag{10}$$

for $p = 0$, 1, and 2. The level-2 probability model assumes the random effects u_{pi} to be normal in distribution with coefficients for linear and quadratic age fixed. The curves in Figure 3 apply the results to the inverse logit transform

$$\hat{\mu}_{ti} = \frac{1}{1 + \exp\{-\hat{\eta}_{ti}\}} \tag{11}$$

that is, the predicted probability of committing serious theft for each of four groups (majority males, majority females, minority males, and minority females).

While Figure 3 is interesting, it is difficult to interpret the polynomial coefficients. More interesting would be parameters that map onto a developmental theory linked to literature from the age-crime curve. SW Raudenbush (2001) transformed the polynomial model for the individual trajectory into a model with three parameters: (*a*) the peak age of offending, that is, the age at which the expected probability of offending is at a maximum; (*b*) the peak probability of offending, that is, the probability of offending at the peak age; and (*c*) the persistence of offending, that is, the extent to which a person continues to offend rather than to desist from offending during the transition to young adulthood. The graph suggests, for example, that females "peak" earlier than males and that, when they do, their offending rate is much lower than that of males. It also suggests that minority and majority youth have quite similar curves except that minority males tend to peak later and to persist. Consistent with the results of Horney et al (1995), we might expect that this persistence reflects the fact that minority males are having a harder time getting jobs and are more likely to remain single than are majority males, although we cannot test this hypothesis with the current data.

The example shows that the level-1 structural model can be recast such that the parameters capture theoretically interesting properties of development. At the same time, the probabilistic part of the model is consistent with the binary nature of the data. A parallel analysis using the frequency of crime reached similar results. In this case, μ_{ti} was the event rate for person *i* at time *t*, the link function was the log link, that is $\eta_{ti} = \log(\mu_{ti})$, and the structural model for η_{ti} is the same as in Equation 9.

Expanding the Class of Level-2 Models

Thus far I have emphasized the importance of "getting the model for individual development right." It is equally important to make sound choices about the model for individual variation in development.

[1]The vast majority of those sampled in the National Youth Survey were either European American or African American. Sample sizes of other subgroups were too small for analysis, especially in light of the low frequency of serious theft.

Simultaneous Equation Models Applications of hierarchical models to date have nearly always specified multivariate regression models at level 2. In terms of our paradigm example of Equations 1 and 2, this means expanding the level-2 model, that is, in an especially simple case,

$$\begin{aligned}\pi_{0i} &= \beta_{00} + \beta_{01}X_i + u_{0i} \\ \pi_{1i} &= \beta_{10} + \beta_{11}X_i + u_{1i} \\ &\textit{with} \quad \begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix} \right].\end{aligned} \tag{12}$$

Here the level-1 parameters of individual change, π_{0i} and π_{1i} , become correlated outcomes predicted by person characteristic X_i. Raudenbush & Sampson (1999) show, however, how to specify and estimate simultaneous equation models in which the random coefficient, say π_{0i}, becomes a mediating variable. In the model below, X_i has both a direct effect on the rate of change π_{1i}, given the intercept π_{0i}, and an indirect effect operating through the intercept. The model simple moves π_{0i} in Equation 13 to the right side of the equation for π_{1i}:

$$\begin{aligned}E(\pi_{1i} \mid \pi_{0i}) &= \alpha_{10} + \alpha_{11}X_i + \alpha_{12}\pi_{0i} \\ &= \beta_{10} - \alpha_{12}\beta_{00} + (\beta_{11} - \alpha_{12}\beta_{01})X_i + \alpha_{12}u_{0i}.\end{aligned} \tag{13}$$

Here $\alpha_{11} = \beta_{11} - \alpha_{12}\beta_{01}$ is the direct effect of X on π_1; $\alpha_{12}\beta_{01}$ is the indirect effect of X on π_1 operating through π_0 and $\alpha_{12} = \tau_{10}/\tau_{00}$.

A Multinomial Prior Assumption In many studies of growth, it is reasonable to assume that all participants are growing according to some common function but that the growth parameters vary in magnitude. For example, vocabulary growth curves for a normative sample of children invariably show upward curvature (acceleration) during the second year of life, and the interesting question about individual differences is the rate of acceleration (Huttenlocher et al 1991). For many other repeated-measures studies, however, the situation is quite different.

Consider a study of changes in depression. It makes no sense to assume that everyone is increasing (or decreasing) with respect to depression. In a normative sample, many persons will never be high in depression, whereas others will always be high; some persons will be recovering from serious depression, but others will become increasingly depressed, and perhaps another group will oscillate between high and low levels of depression. Such "depression curves" can certainly be represented by a polynomial of sufficiently high degree, say three (i.e. the cubic degree) in the level-1 model. However, linear models for the polynomial coefficients, π, at level 2 may not capture the qualitatively different types of trajectories found in the population.

To model these kinds of data, Nagin (1999) has developed a two-level model in which the first level is similar to those discussed in this paper (see also Muthen

2001). However, the second level of the model is reconceptualized such that the population is viewed as (approximately) falling into a fixed number of groups, where each group's development is characterized by a common set of change parameters (πs). The summary of evidence from this model is a set of conditional probabilities for each person: the probability that a person is in group 1, the probability that the person is in group 2, etc. In the example of a study of depression, group 1 might be the "always depressed" group while group 2 might be those who are "becoming depressed." A multinomial regression model then can predict the probabilities of group membership. It may be, for example, that the predictors of being in the "always depressed" group are quite different from the predictors of being in the "becoming depressed" group. This model thus seems especially useful when trajectories of change involve sets of parameters that mark qualitatively different kinds of development. Nagin shows how to test the appropriateness of the assumed number of groups and thus to test alternative models for types of change.

Alternative Probabilistic Models The vast majority of level-2 models assume that departures of the change parameters (the πs) from their predicted values are multivariate normal. This is certainly a convenient assumption, as mentioned, but it may poorly fit the data, and results may not be robust for departures from it. In particular, outlier values of the growth parameters may be far more influential than one would desire, particularly in small-sample settings.

To develop more robust estimation, Seltzer (1993) adopted a multivariate *t*-prior distribution for the random effects (e.g. u_{0i}, u_{1i} in Equation 2). The *t*-prior anticipates more outliers than does the normal prior, and, as a result, model estimates are more resistant to the influence of such outliers. Seltzer embedded this approach within a Bayesian framework—a three-level hierarchical model with the first two levels of the type given by Equations 1 and 2 and a third level specifying a prior for the level-2 parameters. Thum (1997) also adopted a *t*-prior in his multilevel-multivariate model with estimation via maximum likelihood. This model allows covariance structure analysis at each of two levels based on incomplete data.

Causal Inference

Let us consider the problem of causal inference in studies in which the "outcome" is an entire trajectory of change. However, I first review some basic ideas about causation in cross-sectional studies.

Causal Effects in Cross-Sectional Studies What is a causal effect? Many discussions of causal inference and research design neglect to confront this issue. However, a theory that has come to dominate modern thinking in statistics about cause begins with this fundamental question. Pioneered by Rubin (1974) and Rosenbaum & Rubin (1983) and elaborated by Holland (1986), this theory has come to be known as the "Rubin-Rosenbaum-Holland" (RRH). To describe this theory, the simplest case will suffice; given a causal variable ("the treatment") with two

possible values ("experimental" and "control"), for clarity, consider a case in which a child receives either the new experimental approach to daycare (E) or the currently available approach (C) and the outcome, Y, is a measure of self-regulation. If the child receives E, we observe $Y_i(E)$, the outcome of child i under E. However, if that same child receives C, we observe $Y_i(C)$, the outcome of child i under C. The causal effect of the experimental treatment (relative to the control) is defined as the difference between these two potential outcomes,

$$\Delta_i = Y_i(E) - Y_i(C). \tag{14}$$

Several conclusions follow from this definition.

First, the causal effect Δ_i is defined uniquely for each child. The impact of the treatment can thus vary from child to child. Modern thinking about cause thus rejects the conventional assumption that a new treatment adds a constant effect for every child. This assumption, never realistic to scientists or practitioners, was imposed historically to simplify statistical analysis.

Second, the causal effect cannot be observed. If a given child is assigned to E, we will observe $Y_i(E)$, but not $Y_i(C)$. On the other hand, if the child is assigned to C, we will observe $Y_i(C)$ but not $Y_i(E)$. The outcome that cannot be observed is the counter-factual outcome.

Third, it must be reasonable at least to imagine a scenario in which that child could have received either E or C. If it is not possible to conceive of each child's response under each treatment, it is not possible to define a causal effect. Therefore, in current thinking about cause in statistical science, a fixed attribute of a child (say sex or ethnic background) cannot typically be a cause. We cannot realistically imagine how a girl would have responded if she had been a boy or how a black child would have responded if that child had been white. Epidemiologists have referred to such attributes as fixed markers (Kraemer et al 1996), unchangeable attributes that may be statistically related to an outcome but cannot be causes.

Randomized Studies According to RRH, the problem of causal inference is a problem of missing data. If both potential outcomes $Y_i(E)$ and $Y_i(C)$ were observed, the causal effect could simply be calculated for each participant. However, one of the potential outcomes (the counterfactual) is always missing. Randomized studies ensure that the missing counterfactual is missing completely at random. In effect, the mechanism of random assignment ensures that the decision about which outcome is observed [$Y_i(E)$ or $Y_i(C)$] is decided by chance alone. While it is not possible to estimate the causal effect for each participant, a randomized experiment enables unbiased estimation of the average causal effect, that is

$$\Delta = E[Y_i(E) - Y_i(C)], \tag{15}$$

the population average difference between potential outcomes. The estimator is simply the difference between sample means of the experimental and control groups.

Nonrandomized Studies Without random assignment, we cannot assume that the counterfactual is missing at random. First, those selected into the experimental group may be more (or less) disadvantaged than those selected into the control group. If so, the potential outcomes of those assigned to the experimental group would be higher (or lower), on average, than of those assigned to the control group in the absence of a treatment effect. Researchers often attempt to control for confounding variables—preexisting variables that predict treatment group membership and are related to the potential outcomes—by means of matching or statistical adjustments such as the analysis of covariance (Cook & Campbell 1979). However, one can never be sure that the relevant confounding variables have been identified or that the method of adjustment has been completely effective. Second, participants may have information about the likely benefit of choosing to participate in the treatment, information not available to the researcher. For example, a person choosing not to participate in a job-training program may know that she is about to be offered a good-paying job, better than can be expected to result from participating in the program. Such biases cannot be controlled because, by definition, the information needed to adjust for bias is unknown to the researcher. However, certain "natural experiments" can be exploited to avoid such bias. For example, a job-training program may be available only in certain states, ensuring that the choice to participate is not entirely under the control of the participants. Instrumental variable methods (cf Angrist et al 1996, Little & Yau 1998) can be used to exploit such natural experiments in order to minimize bias in nonexperimental studies.

It has often been argued that collecting longitudinal data can also increase the credibility of causal inferences when randomization is impossible. We consider this issue, among others, in the following sections.

Causal Inference in Longitudinal Studies: Between-Subject Causes

Consider a case in which participants are assigned to one of two treatments to assess how the treatment affects an entire trajectory. Now we have a trajectory characterized by change parameters $\pi_i(E)$ if person i is assigned to E and $\pi_i(C)$ if person i is assigned to C. Only one of the two trajectories will be observed, leaving the counterfactual trajectory missing.

Randomized Experiments Under random assignment, the counterfactual trajectory is missing completely at random, making it easy to calculate an unbiased estimate of the average treatment effect as the mean trajectory in the experimental group minus the mean trajectory in the control group—assuming no attrition. If data are missing through attrition, for example, because some participants miss some appointments to be assessed or some drop out of the study, there are three possibilities (Little & Rubin 1987, Schafer 1997).

1. The attrition may lead to data that are missing completely at random (MCAR). Data that are MCAR result when the missing time points are a random sample of all time points or the dropouts are a random sample of all participants. In this case unbiased estimation follows easily using conventional methods, for example analyzing only those cases with complete data. However, to assume that data are MCAR is generally implausible, impossible to verify, and risky.
2. The data may be missing at random (MAR). MAR occurs when the probability of missing a time point is independent of the missing data, given the observed data. This assumption is reasonable when the observed data capture key confounding influences, for example, variables that predict both attrition and the outcome of interest. Under the MAR case, estimation of the treatment effect will be unbiased if (*a*) all of the data are used in the analysis and (*b*) a fully efficient estimation procedure is used. For example, maximum likelihood estimation of hierarchical models (also termed multilevel models of mixed models—see earlier discussion) will efficiently use all of the available time points to estimate the model, assuring unbiased estimation of treatment effects under the MAR case. Under these conditions, the mechanism that produces the "missingness" is "ignorable." Use of multiple, model-based imputation (Little & Rubin 1987, Schafer 1997) will also ensure ignorable missingness when data are MAR.
3. "Nonignorable" missingness arises when the data are neither MCAR or MAR. In this case, the probability of attrition does depend on the missing value, even after controlling for all observed data. Results will be robust to nonignorable missingness to the extent that (*a*) all of the data are efficiently used and (*b*) the fraction of missing information is small.[2] Little (1995) and Hedekker & Gibbons (1997) describe methods for nonignorably missing longitudinal data under the rubric of pattern-mixture models.

Nonrandomized Studies Without random assignment, estimation of treatment effects will, in general, be biased even if there is no attrition. We therefore consider only the "no-attrition case," keeping in mind that the concerns about attrition that apply in randomized experiments will also apply in nonrandomized experiments.

Below I consider how longitudinal studies can overcome or reduce bias when it is impossible to randomly assign participants to treatments. I consider studies with two time points, three time points, and more than three time points.

[2]The fraction of missing information is not the fraction of missing data points. It is, rather, related to the amount of variation in missing data that is not explained or accounted for by observed data. If the associations between the observed and missing data are strong, the fraction of missing data will be small even if the fraction of missing cases appears quite large.

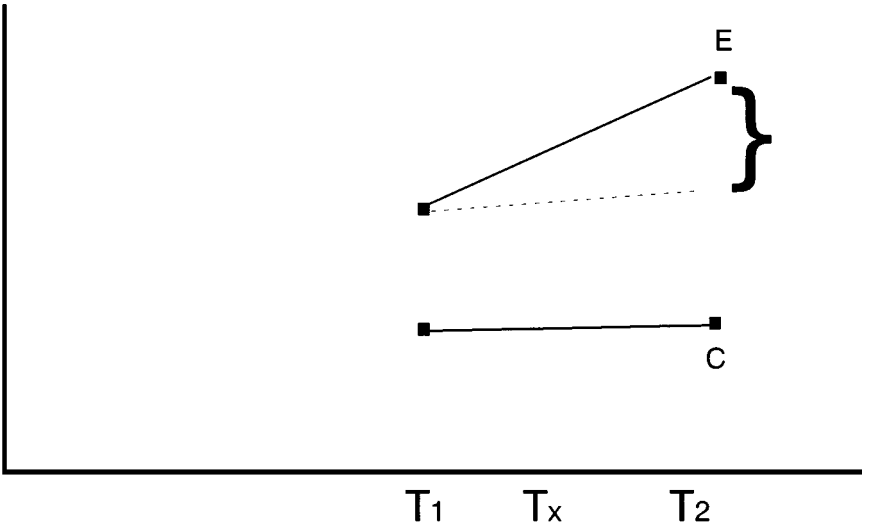


Figure 4 Causal inference when children are growing: pre-post design.

Causal Comparative Studies with Two Time Points Consider a nonrandomized cross-sectional study. As mentioned above, the mean difference between outcomes of experimental and control groups is not, in general, an unbiased estimate of the treatment effect. One strategy to cope with bias is to add a pretest. Assuming Y is measured on a common scale at each time point, the difference in gain scores is arguably less biased than the difference in post-test scores. Consider, for example, Figure 4. The figure shows the average pre-post trajectory for experimental groups and for controls. The length of the interval in curly braces is presumably the treatment effect. Using Equation 12 as a model, this estimate is

$$\hat{\beta}_{11}\,(T_2 - T_1), \tag{16}$$

that is, the difference in mean growth rates between the two groups (β_{11}) multiplied by the time elapsed between the pre- and post-test.

The problem is that such an estimate assumes that the two groups would have experienced identical growth rates in the absence of the treatment. Given nonrandom assignment, this seems improbable. A design having two time points gives no information about the rate of growth of the experimental group in the absence of a treatment. This fact is illustrated in Figures 5*A* and 5*B*. Figure 5*A* shows parallel growth of the two groups before the treatment, making the interval in curly braces plausible as a measure of treatment impact. However, Figure 5*B* shows nonparallel growth before the implementation of the treatment, undermining any case for that estimate. Two time points provide no basis for adjudicating

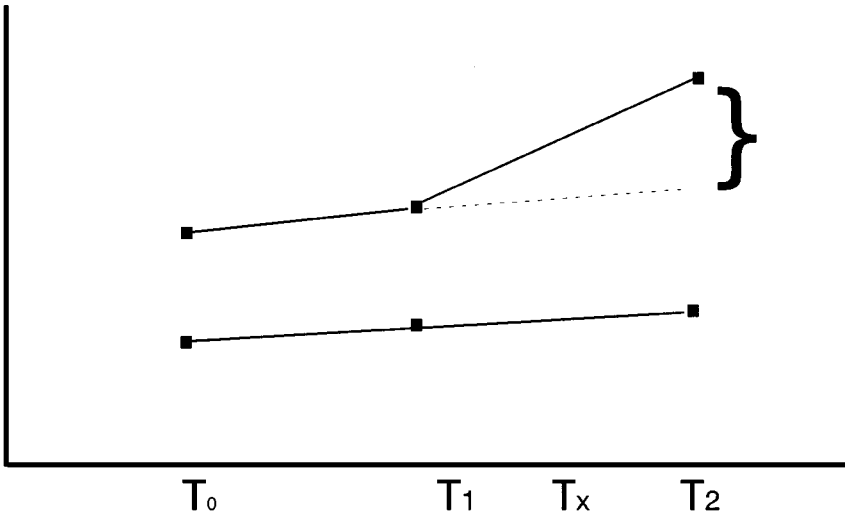


Figure 5*A* Scenario 1: parallel growth in absence of treatment.

between Figures 5*A* and 5*B* or any other pretreatment growth configuration. The problem does not follow from the use of gain scores. Bryk & Weisberg (1977) showed that all strategies of linear adjustment (including not only the gain score adjustment but also the analysis of covariance) are subject to biases that cannot be investigated with only two time points.

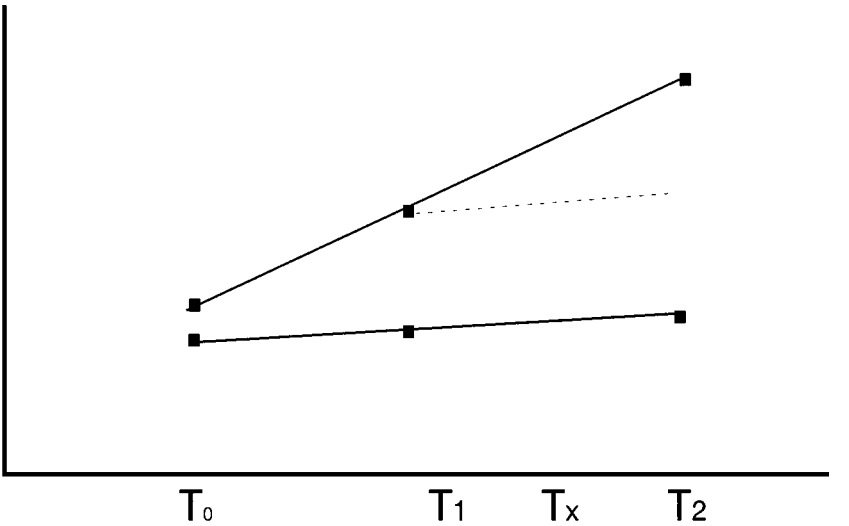


Figure 5*B* Scenario 2: Fan spread growth in absence of treatment.

Three Time Points Adding a third time point, in particular, a "pre-pre-test," adds substantially to the information available in estimating the treatment effect. When we elaborate Equations 1 and 2 to incorporate this information, our level-1 model becomes

$$Y_{ti} = \pi_{0i} + \pi_{1i} a_{ti} + \delta_i Z_{ti} + e_{ti}. \tag{17}$$

Here Z_{ti} is a dummy variable taking on a value of 1 at the third time point and 0 otherwise, so that δ_i is the deflection experienced by participant i between times 2 and 3, that is, when the treatment is implemented. The other terms in Equation 4 are defined as in Equation 1. We now elaborate the level-2 model to allow the deflection to depend on group membership:

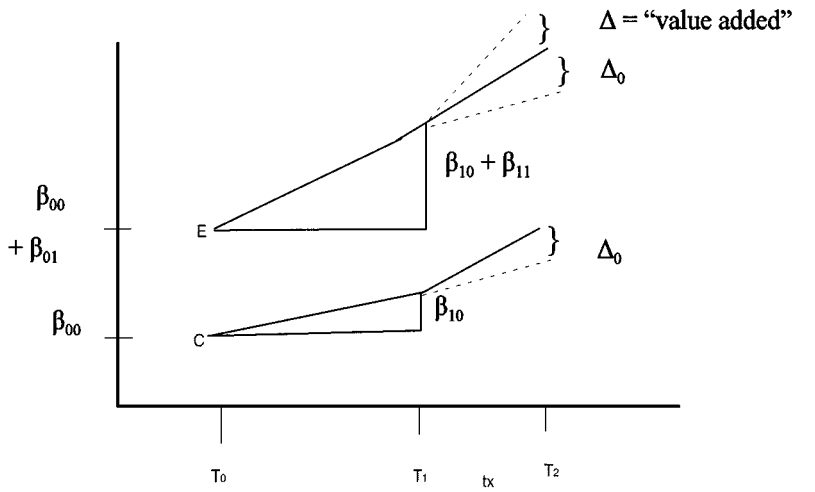
$$\begin{aligned} \pi_{0i} &= \beta_{00} + \beta_{01} X_i + u_{0i} \\ \pi_{1i} &= \beta_{10} + \beta_{11} X_i + u_{1i} \\ \delta_i &= \Delta_c + \Delta X_i. \end{aligned} \tag{18}$$

As displayed in Figure 6, this model allows estimation of (*a*) different mean initial status for *C* and *E*; (*b*) different pretreatment growth rates for *C* and *E*; and (*c*) different nonlinear deflections for *C* and *E* at time of treatment. The difference between these nonlinear deflections is the presumptive treatment effect. Clearly, adding a single time point has added a substantial amount of information relevant to estimating the treatment effect.

More Than Two Time Points The quasi experimental design with three points, while much stronger than the design with two points, may be criticized for assuming that both groups would have the same nonlinear deflection in the absence of a treatment. Adding a fourth time point can distinguish nonlinear deflections and a treatment effect. Adding a fourth time point (a second post-test) can also enable the estimation of a randomly varying effect of treatment. This brings the model into closer alignment with the notion of participant-specific effects that is central to the potential-outcomes framework.

Time-Varying Treatments

We have limited the discussion so far to the case in which participants experience one and only one treatment. Yet many interesting causal questions involve time-varying treatments. Every time a child speaks, a caregiver's response can be conceived as a treatment. A time series unfolds in which each speech event triggers such a "treatment," which affects the next speech event, and so on. Other examples include the use of medical treatment: a physician will assess a child and prescribe (or not prescribe) a drug; the physician will then follow up later with another assessment and another decision to prescribe or not, etc. In school, children are assigned to special education (or not) based on assessments of academic progress



Model for individual growth

$$Y_{it} = \pi_{0i} + \pi_{1i}\alpha_{ti} + \delta_i D_{2i} + e_{ti}$$

Model for population variation in growth curve

$$\pi_{0i} = \beta_{00} + \beta_{01}X_i + U_{0i}$$

$$\pi_{1i} = \beta_{10} + \beta_{11}X_i + U_{1i}$$

$$\delta_i = \Delta_0 + \Delta X_i$$

$$X_i = \{1 \text{ if E}, 0 \text{ if C}\}$$

Figure 6 Estimating the treatment effect based on three time points.

and behavior and are then "mainstreamed" (or not) based on subsequent behavior, which has arisen, in part, as a result of the first placement decision.

In all of these examples, the assessment, treatment, later assessment, and later treatment are interconnected. There is a serious risk in these examples that the analyst will estimate the effect of treatment on the outcome with bias. Children who are assigned medication will do worse than those who are not, making the medication look harmful; children assigned to special education will look worse

than those who are not, making special-education placement look harmful; and the mainstreamed will look better than those who are not, making mainstreaming look helpful—even if the medication, placement, and mainstreaming have no effect.

A dynamic-treatment regime arises when a time-varying treatment or "dose" is calibrated to the current status of the participant. It is normally true that "the worse the current status, the larger the dose." The "regime" is the set of rules that assign doses to children by their statuses. Methodological research shows that one can make sound inferences about regimes, for example, by assigning participants at random to the regime (Robins et al 2000). However, if a regime is strictly adhered to, it is not possible, within regimes, to assess the causal effect of a dose. Only if the dose can vary, given current status, is it possible to make a causal inference about the dose.

The problem of causal inference for time-varying treatments, assuming no randomization, is extraordinarily challenging. It is a cutting-edge issue on which progress is being made, but methodological workers in this area have not yet achieved a consensus on the best ways to proceed.

Summary and Conclusions

A statistical model is essential in precisely defining each participant's trajectory and facilitating comparisons of persons by comparing their trajectories. Based on such a model, one can define and estimate the reliability of measures of change. One can also define effect sizes and, with supplementary data at hand, compute the power of alternative designs (Hedeker & Gibbons 1997; Muthen & Curran 1997). Alternative models posit alternative views of confounding variables and alternative ways in which attrition might affect the precision and bias of inferences.

An important new line of research considers how to extend modern thinking about causal effects to longitudinal data. According to this modern theory, a causal effect is person specific; it is the difference between how a person would respond under one condition and how that same person would respond under a different condition. The causal effect can thus vary among people. However, it cannot be directly computed, because invariably, a person will experience one condition but not the other at any given time. The set of responses that might occur for a person under the varied conditions possible is called the set of potential outcomes. A confounder is a pretreatment characteristic of a participant that is related to both the propensity to receive the treatments and to the potential outcomes. Valid causal inferences require that the propensity to receive treatments is independent of the potential outcomes, given the potential confounders on which the investigator has collected data. Randomized experiments ensure independence. In nonrandomized experiments, knowledge of predictors of propensity to receive the various treatments is essential in seeking valid causal inferences. In the nonrandomized setting, the burden of proof is on the shoulders of the investigator to show that the relevant confounders have been taken into account, and it will be difficult to avoid lingering disagreement about the success of the attempt to take these into account.

It is often essential in developmental research to recognize that participants are naturally growing or changing even in the absence of a treatment. A model for individual growth in the absence of treatments can be extremely useful in identifying causal effects, especially in the challenging nonrandomized setting. Multiple time points can lay a basis for interrupted time series designs with improved validity of causal inference in nonrandomized studies.

When the growth of interest is psychological, it is challenging to define clearly the dimensions in which children are growing, to devise assessments that are sensitive to growth, and to evaluate the capacity of alternative designs to reliably gauge individual differences in growth. Given that repeated measures are required to assess growth, the problem of attrition cannot be ignored. Attrition can not only weaken statistical precision but also introduce bias, even in randomized experiments. In these settings, it is incumbent on the investigator to study the predictors of attrition and to devise ways to control attrition in assessing effects.

Finally, time-varying treatments pose special challenges because a treatment will often be tailored to the past behavior of the participants. Valid causal inference in this setting constitutes an area of intense interest in methodological research.

ACKNOWLEDGMENTS

Research reported here was supported by grants to the Project on Human Development in Chicago Neighborhoods by the MacArthur Foundation, the National Institute of Justice, and the National Institute of Mental Health.

LITERATURE CITED

Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91(434):444–72

Blalock H. 1969. *Theory Construction*. Englewood Cliffs, NJ: Prentice-Hall

Blau P, Duncan O. 1967. *The American Occupational Structure*. New York: Wiley & Sons

Bryk A, Raudenbush S. 1992. *Hierarchical Linear Models for Social and Behavioral Research: Applications and Data Analysis Methods*. Newbury Park, CA: Sage

Bryk A, Weisberg H. 1977. Use of the nonequivalent control gorup design when subjects are growing. *Psychol. Bull.* 84(5):950–62

Carter L. 1984. The sustaining effects study of compensatory and elementary education. *Educ. Res.* 13(7):4–13

Collins L. 2001. Measurement reliability for static and dynamic categorical latent variables. In *New Methods for the Analysis of Change*, ed. L Collins A Sayer. Washington, DC: Am. Psychol. Assoc. In press

Collins L, Wugalter S. 1992. Latent class models for stage-sequential dynamic latent variables. *Multivar. Behav. Res.* 27:131–57

Cook T, Campbell D. 1979. *Quasi-Experimentation*. New York: Rand McNally

Cox D, Oakes D. 1984. *Analysis of Survival Data*. London: Chapman & Hall

Diggle P, Liang K, Zeger S. 1994. *Analysis of*

Longitudinal Data. New York: Oxford Univ. Press

Duncan O. 1966. Path analysis: sociological examples. *Am. J. Sociol.* 72(1):1–16

Duncan O, Featherman D, Duncan B. 1972. *Socioeconomic Background Achievement*. New York: Seminar Press

Francis DJ, Fletcher JM, Stubting KK, Davidson KC, Thompson NM. 1991. Analysis of change: modeling individual growth. *J. Consult. Clin. Psychol.* 39(1):27–37

Goldstein H. 1995. *Multilevel Statistical Models*. New York: Wiley & Sons. 2nd ed.

Gottfredson M, Hirschi T. 1990. *A General Theory of Crime*. Stanford, CA: Stanford Univ. Press

Gottman J. 1995. *The Analysis of Change*. Hillsdale, NJ: Erlbaum

Hedeker D, Gibbons R. 1997. Application of random effects pattern mixture models for missing data in social sciences. *Psychol. Methods* 2(1):64–78

Holland P. 1986. Statistics and causal inference. *J. Am. Stat. Assoc.* 81(396):945–60

Horney J, Osgood D, Marshall I. 1995. Criminal careers in the short-term: intra-individual variability in crime and its relation to local life circumstances. *Am. Sociol. Rev.* 60: 655–73

Huttenlocher J, Haight W, Bryk A, Seltzer M. 1991. Early vocabulary growth: relation to language input and gender. *Dev. Psychol.* 27(2):236–49

Joreskog K, Sorbom D. 1979. *Advance in Factor Analysis and Structural Equation Models.* Cambridge, MA: ABT Assoc.

Kalbfleisch J, Prentice R. 1980. *The Statistical Analysis of Failure Time Data*. New York: Wiley & Sons

Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ. 1996. Coming to terms with the terms of risk. *Arch. Gen. Psychiatry*

Laird N, Ware J. 1982. Random-effects models for longitudinal data. *Biometrika* 65(1):581–90

Little R, Rubin D. 1987. *Statistical Analysis with Missing Data*. New York: Wiley & Sons

Little R, Yau L. 1998. Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. *Psychol. Methods* 3(2):147–59

Little R. 1995. Modeling the drop-out mechanism in repeated measures studies. *J. Am. Stat. Assoc.* 90:1112–21

Longford N. 1993. *Random Coefficient Models.* Oxford, UK: Clarendon

McCullagh P, Nelder J. 1989. *Generalized Linear Models*. London: Chapman & Hall. 2nd ed.

Meredith W, Tisak J. 1990. Latent curve analysis. *Psychometrika* 55:107–22

Moffitt T. 1993. Adolescence-limited and life-course persistent antisocial behavior: a developmental taxonomy. *Psychol. Rev.* 100:674–701

Muthen B. 2001. Second generation structural equation modeling with a combination of categorical and continuous latent variables: new opportunities for latent class-latent growth modeling. In *New Methods of the Analysis of Change*, ed. L Collins, A Sayer. Washington, DC: Am. Psychol. Assoc. In press

Muthen BO, Curran PJ. 1997. General longitudinal modeling of individual differences in experimental designs: a latent variable framework for analysis and power estimation. *Psychol. Methods* 2(4):371–402

Nagin DS. 1999. Analyzing developmental trajectories: a semi-parametric, group-based approach. *Psychol. Methods* 4(2):139–57

Raudenbush SW. 2001. Toward a coherent framework for comparing trajectories of individual change. In *Best Methods for Studying Change*, ed. L Collins, A Sayer. Washington, DC: Am. Psychol. Assoc. In press

Raudenbush SW, Chan WS. 1993. Application of a hierarchical linear model to the study of deviance in an overlapping cohort design. *J. Consult. Clin. Psychol.* 61(6):941–51

Raudenbush SW, Sampson R. 1999. Assessing direct and indirect associations in multilevel designs with latent variables. *Sociol. Methods Res.* 28(2):123–53

Robins JM, Greenland S, Hu F-C. 2000. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcomes. *J. Am. Stat. Assoc.* In press

Rogosa D, Brandt D, Zimowski M. 1982. A growth curve approach to the measurement of change. *Psychol. Bull.* 90:726–48

Rosenbaum P, Rubin D. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 17:41–55

Rubin D. 1974. Estimating causal effects of treatments in randomized and non-randomized studies. *J. Educ. Psychol.* 66:688–701

Schafer J. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall

Seltzer M. 1993. Sensitivity analysis for fixed effects in the hierarchical model: a Gibbs sampling approach. 18(3):207–35

Seltzer MH, Frank KA, Bryk AS. 1994. The metric matters: the sensitivity of conclusions concerning growth in student achievement to choice of metric. *Educ. Eval. Policy Anal.* 16(1):41–49

Sewell W, Hauser R. 1975. *Education, Occupations, and Earnings: Achievement in the Early Career*. New York: Academic

Singer J, Willett J. 1991. Modeling the days of our lives: using survival analysis when designing and analyzing longitudinal studies of duration and the time of events. *Psychol. Bull.* 110(2):268–90

Thum Y. 1997. Hierarchical linear models for multivariate outcomes. *J. Educ. Behav. Stat.* 22:77–108

Annual Review of Psychology
Volume 52, 2001

CONTENTS