J. Child Psychol. Psychiat. Vol. 42, No. 1, pp. 141–162, 2001
Cambridge University Press
© 2001 Association for Child Psychology and Psychiatry
Printed in Great Britain. All rights reserved
0021–9630/01 $15.00+0.00

Multilevel Modelling of Hierarchical Data in Developmental Studies

Michael H. Boyle
McMaster University and Hamilton Health Sciences Corporation, Hamilton, Canada

J. Douglas Willms
University of New Brunswick, Fredericton, Canada

This report attempts to give nontechnical readers some insight into how a multilevel modelling framework can be used in longitudinal studies to assess contextual influences on child development when study samples arise from naturally formed groupings. We hope to achieve this objective by: (1) discussing the types of variables and research designs used for collecting developmental data; (2) presenting the methods and data requirements associated with two statistical approaches to developmental data—growth curve modelling and discrete-time survival analysis; (3) describing the multilevel extensions of these approaches, which can be used when the study of development includes intact clusters or naturally formed groupings; (4) demonstrating the flexibility of these two approaches for addressing a variety of research questions; and (5) placing the multilevel framework developed in this report in the context of some important issues, alternative approaches, and recent developments. We hope that readers new to these methods are able to visualize the possibility of using them to advance their work.

Keywords: Longitudinal studies, development, growth curve analysis, survival analysis, multilevel models.

Abbreviations: CBT: treated with cognitive behavioural therapy; MED: treated with anti-depressive medication; OLS: ordinary least squares.

The study of contextual influences on child development has witnessed remarkable expansion in the past few years. This expansion has been stimulated by three activities. One, researchers have been thinking about new ways of conceptualizing and measuring broader contextual forces embedded within intentionally formed human groupings such as neighbourhoods, families, and schools. The operationalization of contextual forces in different environments is creating a new array of variables for developmentalists to study in the nature-nurture continuum (Boyce et al., 1998).

Two, national governments are investing substantially in large-scale population-based studies of child development, and facilitating the analysis of the resulting data by interested researchers. For example, the government of Canada launched the National Longitudinal Study of Children and Youth, which began in 1994 with a probability sample of over 20,000 children aged 0 to 11 years, and included data collection from parents, teachers, and school administrators. These children and their families are being followed longitudinally, with data collected biennially (Special Surveys Division, 1996). This year, it is launching its Youth-in-Transition Study, which will follow cohorts of 15- and 18-year-old youth annually for at least 3 years. Studies like these are sampling naturally formed groupings (families, neighbourhoods, and schools) to examine individual and contextual influences on child development. The resulting data structures are complex and inherently hierarchical: the shared affiliations arising within each level results in correlation or nonindependence among measured responses.

Three, statistical methods and computer software have been developed to overcome the analytical dilemmas associated with correlated measurement (see reviews by Horton & Lipsitz, 1999; Kreft, De Leeuw, & Van der Leeden, 1994; Zhou, Perkins, & Hui, 1999). Correlated measurement arises from statistical dependence among observations. A requirement for the use of conventional statistical methods is that the observations must be independent; for example, the observations of any one individual cannot be systematically related to the observations of any other individual. Statistical dependence is usually present when individuals are sampled from naturally formed clusters (e.g., measurements on siblings within families, or families within neighbourhoods), and always present when assessments are repeated on the same individual. The statistical methods presented in this report go by a variety of names—random effects models, empirical Bayes models, and hierarchical linear models or multilevel models (Gibbons et al., 1993). They are creating important opportunities to answer complex questions about influences on development, including those that arise from shared experiences in natural groupings.

Requests for reprints to: Michael H. Boyle, Centre for Studies for Children at Risk, McMaster University Faculty of Health Sciences and Hamilton Health Sciences Corporation, Patterson Building, Chedoke Division, Hamilton Health Sciences Corporation, Box 2000, Hamilton, Ontario, Canada L8N 3Z5
(E-mail: boylem@fhs.csu.mcmaster.ca).

The objective of this report is to show how multilevel modelling can be used to analyze child development as a continuous or transitional process when study samples arise from naturally formed groupings. We hope to achieve this objective in the following way: (1) by discussing the types of variables and research designs used for collecting developmental data; (2) by presenting the methods and data requirements associated with two statistical approaches to developmental data—growth curve modelling and discrete-time survival analysis; (3) by describing the multilevel extensions of these approaches, which can be used when the study of development includes intact clusters or naturally formed groupings; (4) by demonstrating the flexibility of these two approaches for addressing a variety of research questions; and (5) by placing the multilevel framework developed in this report in the context of some important issues, alternative approaches, and recent developments.

Before proceeding, we acknowledge that this report was written by an epidemiologist with modest statistical assets (MB) under the watchful eye of an analyst with experience in applying these models (JDW). In a sense, the first author is part of the target audience: the division of responsibilities was intended to result in a paper that would be sensitive to the needs and comprehension levels of nontechnical readers without incurring the wrath of the more mathematically inclined. In the last 10 years numerous texts and journal articles on the topics addressed in this paper have appeared, and many of these are cited throughout. However, a special acknowledgement goes to the work of Judy Singer and John Willett, who have a gift for taming the statistically complex and making it accessible to the uninitiated. We recommend their work wholeheartedly (e.g., see Willett, Singer, & Martin, 1998).

Central Concepts

The title, *Multilevel Modelling of Hierarchical Data in Developmental Studies* encompasses several key concepts. Developmental studies are investigations focused on the evolution of individual characteristics and experiences. Individual characteristics can be any human feature—for example, biological, psychological, social—that are both measurable and subject to change. Experiences are events or phenomena that may or may not happen to individuals over a defined period, which can gradually shape a person's life course, or can mark a transition or turning point. There are numerous ways to calibrate the evolution or time course of developmental phenomena—minutes, hours, days, months, years—and the selection of this unit depends on the nature of the inquiry.

Hierarchical data are numerical summaries obtained from measured variables that exhibit a specific relational structure. This relational structure can be viewed as a series of levels from the most specific observation, usually identified as level 1, through an ever-expanding array of levels defined by observations that measure characteristics of naturally formed subsets. The study of developmental phenomena requires measurement of the same concepts assessed in temporal sequence on the same individuals. Each occasion of measurement or wave of data collection in a developmental study contributes to a naturally formed subset of observations that are nested within individuals. Because these observations belong to a tnaturally formed subset, an expectation exists that they will be correlated with one another. The presence of such correlations violates the assumption of independence that applies to traditional statistical methods.

To extend this idea further, it may be helpful to consider the relational quality of hierarchical data from the perspective of shared experience. Repeated assessments are shared experiences linked through an individual. A developmental study of sibships takes into account the shared experiences of individual children linked through families. A developmental study of neighbourhoods takes into account the shared experiences of sibships linked through defined geographical areas. By convention, the lowest level measurement in hierarchical data structures (level 1) is said to be at the micro level (Kreft & De Leeuw, 1998, p. 1). Thus, repeated assessments on individuals taken in developmental studies constitute micro data. All higher-level measurements are said to be at the macro level. The characteristics of individuals participating in developmental studies constitute macro data, existing at level 2. In a developmental study of sibships, the characteristics of participating families would constitute macro data, existing at level 3. This same study could be expanded to encompass neighbourhoods whose characteristics would constitute macro data, existing at level 4. The opportunities for identifying naturally formed subsets of individuals embedded in hierarchical data structures are almost limitless. It is relatively easy to visualize how schools, workplaces, and geographical areas could be used to divide individuals into subsets exposed to shared experiences. In principle, the correlated observations that arise in clusters or naturally formed groups are no different from the correlated observations that come from repeated assessments on the same individual.

The discussion in the previous paragraph identified a data hierarchy with four possible levels: a micro level consisting of repeated assessments (level 1), and macro levels consisting of individual children (level 2), nested within families (level 3), and nested within neighbourhoods (level 4). There are specific conditions when the relational quality of hierarchical data require the use of statistical methods, such as multilevel modelling, that account for within-group correlations. These conditions are dictated by sampling and design. For example, the use of cluster sampling in developmental studies or the random allocation of natural groupings in experiments necessitate the use of special statistical methods to account for within-group correlations. The importance of using these methods is related directly to the extent of response dependency as estimated by the size of the within-group correlations. When clusters are sampled or intact groupings used in experiments, statistical methods that account for within-group correlations are only irrelevant when these correlations are shown to be zero. The relational structure that must be accounted for in hierarchical data flows directly from the sampling and design parameters. For example, a developmental study that begins by randomly selecting neighbourhoods and then randomly selecting children within neighbourhoods would have to account for two types of correlation at the macro level: within-neighbourhood correlations (level 3) and within-individual correlations (level 2). Although some children may come from the same family by chance, it is not necessary to account for this statistically. In other words, the use of special analytical models that account for correlated observations within groups is optional when groupings arise as a by-product of random processes. If we were to change the present example, and families were selected as a means for sampling intact sibships, then statistical methods would have to be used to account for within-family (or sibship) correlations of response. Now, at the macro level there would be three types of within-group correlations to consider: within neighbourhood (level 4), within family (level 3) and within individual (level 2).

Multilevel modelling is a method of analysis used to test the adequacy of mathematical models for summarizing relationships among measured variables assessed within different clusters or groupings that form an hierarchical data structure. Typically, a researcher posits a model describing how the world works, based on some theory, a review of previous research, prior experience, or simply intuition. This is expressed as a mathematical model, or equation, which species the relationships among variables. For example, in answer to the question, "what happens to variable Y as variable X changes?," a researcher could posit a linear relationship between the two variables, and specify the equation of a straight line:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1}$$

where Y_i could be a person's score on some test, and X_i could be the amount of instructional time allocated randomly to 20 individuals. The subscript i simply denotes that we have a separate score for each individual. The line is defined by several parameters, including an intercept, or constant value, β_0, and a regression slope, β_1 which indicates the expected change in Y for a one-unit change in X. The model also allows for a person's score on Y to deviate from the line by an amount, ε_i, referred to as the residual.

The model specified by equation (1) is only a theoretical model. The researcher collects data representing the constructs Y and X, and "fits" the data to the mathematical model. This entails determining the "best" straight line through the data, based on some criteria. The most common technique is ordinary least squares (OLS) regression, which estimates the model parameters β_0 and β_1 by minimizing the squares of the residuals. With the estimates of β_0 and β_1 derived from the model, one can estimate a predicted value for Y for a person with a particular value of X_i, say X_0:

$$\hat{Y}_0 = \beta_0 + \beta_1 X_0 \tag{2}$$

The small "hat" (ˆ) on Y_0 simply denotes that it is a "predicted value". Another way to think about OLS regression then is that it essentially minimizes the differences between people's observed values of Y and the predicted values associated with their observed values on X. The adequacy of the researcher's model is assessed by determining the magnitude of the differences between observed and predicted values.

We suspect that the OLS regression technique will be familiar to most readrs. Some understanding and comfort with regression will render intelligible the discussions to follow, because the statistical approaches described in this report are either variants of OLS regression modified to accommodate a developmental perspective, or logistic regression techniques which differ in the way that the regression parameters are estimated, but follow the same logic. The core difference between regression models of the type shown in equation (1) and multilevel (regression) models to be presented later is the estimation of residual variation or error. In multilevel models, total residual variation or error is partitioned among levels of the data hierarchy. In standard regression, there is only one estimate of residual variation or error, and it is identified with the level of analysis chosen for study.

Two Types of Developmental Data

Most research questions on child development attempt to characterize changes over time in response variables describing individuals' characteristics or experiences. Although it is impossible to catalogue all of the specific questions which developmentalists might pursue, it is possible to characterize most of them according to whether development is conceived of as a *continuous* or *transitional* process. This largely determines the type of data representing the response variable, and bears directly on the mathematical model and statistical approach used to describe relationships among variables.

Development as a Continuous Process

The first process views development as a continuous phenomenon, as depicted in Fig. 1 a. Much of growth and development focuses on the continuous acquisition and loss of functional characteristics, for example, the acquisition of language throughout childhood, or the loss of memory during later adult life. To measure the acquisition of language, a researcher might observe and record the number of words voiced by a child in response to a specific task and then repeat the assessments at monthly intervals. Although the measures are taken cross-sectionally during discrete intervals, the underlying process is both instantaneous and continuous: it can be conceptualized as a smoothly evolving function of time. The research by Huttenlocher and her colleagues is a good example; they collected data for a sample of toddlers at several intervals between 14 and 26 months to assess the rate of children's vocabulary growth (Huttenlocher, Haight, Bryk, & Seltzer, 1988; see also Bryk & Raudenbush, 1992).

This perspective of change in the acquisition and loss of functional characteristics as instantaneous and continuous is directly analogous to the concept of speed. Mapping the acceleration and deceleration of a moving vehicle over time would yield a smooth plot of time on the horizontal or X-axis and distance travelled on the vertical or Y-axis. In theory, there is no disjunction in the movement of a car. It may appear to stop suddenly but only does so in continuous decrements of motion.

Development as a Transitional Process

The second process views development as a transitional phenomenon, as depicted in Fig. 1 b. In addition to studying the evolution of functional characteristics, researchers focus their attention on the occurrence and timing of major life experiences and transition points that have important developmental implications. Some examples include starting school, grade repetition, family dissolution, and pregnancy. To study grade repetition as a developmental outcome, a researcher might assemble a group of children who have not failed, monitor them for a number of years, identify the children who fail during that period, and record when these failures occur. In this context, development is not a smooth function of time for the individual; rather, it is a status change with some probability of occurrence during a period of observation.

The study of grade failure among a group of children

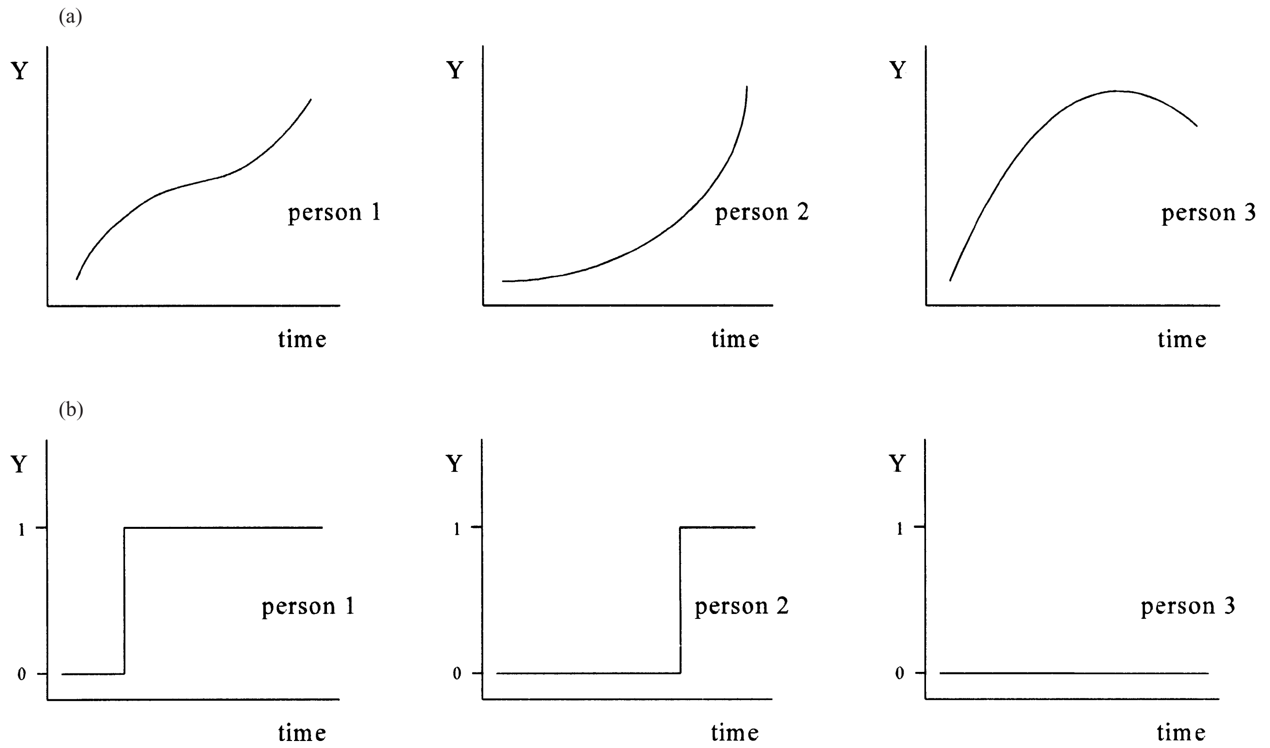

Figure 1. Development as (a) a continuous process; (b) a transitional process.

outlined above represents a single, irreversible transition. One can also study transitions as *competing events*, for example, transitions into work classified as part-time and full-time. One can also envisage *reversible* or two-way transitions, for example, recovery and relapse subsequent to the detection of disease. However, whether the transitional process is irreversible or reversible, or into a single outcome or competing outcomes, an important distinction between transitional and continuous processes is that for transitional processes the "event" may never occur for some individuals (e.g., person 3 in Fig. 1 b) during the course of the study (or even their lifetime). The "loss" of these individuals, called censoring, must be accommodated appropriately in the analysis.

Data Requirements

In the previous section we indicated that statistical approaches to developmental data are based on "measurement of the same concepts assessed in temporal sequence on the same individuals". Thus, longitudinal data are a fundamental requirement for developmental studies. The type of response variable has an important bearing on the selection of an appropriate statistical approach. This is determined by the expected shape of the frequency distribution for the response variable, or, more formally, its theoretical probability distribution.

Studies of continuous processes. In developmental studies of continuous processes, the response variable is usually measured as a continuous phenomenon, as indicated in Fig. 1 a. There are numerous examples of such variables: height, weight, body mass, language acquisition, school readiness, intelligence, aggressive behaviour, and mood, to name a few. To be useful, the scale values associated with continuous measures should: represent interval or ratio level measurement; have identical meaning across the developmental period of study; fluctuate freely without abutting a high or low end point (i.e., are not subject to floor or ceiling effects); and be calibrated in their raw or unstandardized form (Willett & Sayer, 1994). Continuous measures are usually expected to have normal or Gaussian distributions with no restriction on the theoretical range of scores. In regression analysis, response variables with normal distributions are modelled as a linear function of the regression coefficients estimated for the predictor variables. This means that the estimated regression coefficients times the observed values of their corresponding predictor variables, for example, β_1 times X in equation (1), gives a sensible prediction of the response. It also means that for adequate models of the response, the residual terms (observed minus predicted values of the response) are expected also to exhibit a normal distribution.

When response data are ordinal, and the researcher is comfortable in making some assumptions about the magnitude of the intervals between categories, there are reasonably simple techniques for transforming responses into interval data (Mosteller & Tukey, 1977). Although this approach comes recommended, it should not be considered a cure-all for weak measurement, and when used should be tested for robustness under different assumptions regarding the size of the intervals between categories. In the later presentation of multilevel modelling for analyzing child development as a continuous process, we focus attention on normally distributed response data. The multilevel approach is not restricted to these outcomes: discrete response variables can be modelled in this developmental context, and a very brief report on these methods appears in the commentary.

Studies of transitional processes. In developmental studies of transitional processes, the response variable is usually measured as a discrete phenomenon formed from a binary classification coded 0 and 1, as indicated in Fig. 1 b. The response variable could be nominal with several categories, as in the case of competing events (described earlier), or ordinal, such as grades assigned to students by a classroom teacher. Moreover, the transitions could be reversible. In the later presentation of discrete-time survival analysis, we focus attention on dichotomous endpoints and irreversible transitions. The statistical techniques for dealing with competing events or reversible transitions are rather complex, and beyond the scope of this paper.

Covariates. The requirements for the type of data used as independent variables or *covariates*—that is, variables that can potentially explain fluctuations in the response variable—are less stringent. The models we discuss in this paper can handle any type of data for the covariates, although ordinal data with many levels tend to result in cumbersome models. In developmental research, a distinction is made between *time-independent* variables, which have values that do not change over time, and *time-dependent* variables, which have values that can fluctuate with the passage of time. A person's birthweight and their sex are examples of time-independent variables. In developmental research, such variables can be used to predict the response variable or to control for unwanted differences between subjects. Often the interest is in their statistical interaction with other variables in moderating the responses; for example, does a treatment differ in its effect for males and females? Time-dependent variables include the vast array of personal characteristics (e.g., behaviours, feelings, attitudes, and cognition) and contextual influences (e.g., family functioning, neighbourhood cohesion), which are subject to change from one occasion of measurement to the next. These variables have the same role in regression analyses as time-dependent variables, but they also provide an opportunity to study the dynamic interplay over time between alterable characteristics and subject responses.

In developmental research addressing issues of causation, variables may be classified as *exogenous*—arising from a set of forces external to a proposed model, or *endogenous*—not known or assumed to be statistically independent of response and included in the proposed model for theoretical reasons. Although the distinctions between exogenous and endogenous variables can be blurred in practice, the availability of hierarchical data structures expand the research opportunities for distinguishing between these types of variables. It is not unreasonable, for example, to assume that influences of variables measured at higher levels are increasingly independent of variables measured at lower levels.

Research Design

Research design for developmental studies depends largely on the research question, which determines the timing of data collection and the sampling strategy. In principal, we would like to have data on the response variable for each person in a study, at every point in time. We would also like to have measurements at every time point for covariates that change over time. Of course, this is impossible. The aim of research design in developmental studies is to capture data which are sufficient for summarizing and explaining the unfolding of individual characteristics and experiences over time. For example, if the "true" developmental process were similar to one of the curves describing a person in Figs. 1a or 1b, then the aim of research design would be to capture sufficient data so that we could accurately portray that process. Thus, if growth were exponential, as it is for person 2 in Fig. 1a, measurement on three occasions would be insufficient.

This section discusses considerations regarding the measurement schedule—the number of occasions and their temporal pattern, and the sampling strategy. We contend that the minimum requirement is data describing individuals on at least three occasions, but we do not entirely dismiss all data collected in cross-sectional or follow-up (two time-point) studies. We will discuss these first, before elaborating on measurement schedules and sampling strategy.

Cross-sectional studies. Cross-sectional studies are ones in which all of the information is collected on a single occasion. Can such studies address developmental questions? They can in theory by retrieving information from the past, either by accessing available recorded data or by relying on the memories of respondents. Is there a criterion for determining whether or not such studies should be considered developmental? The answer is yes: it must be possible to record different values of the response variable at certain times over a specified interval. For example, in studies of diseases, epidemiologists often sample a group of "cases" (individuals classified with a particular disease) and a separate group of "controls" (individuals without the disease), and then compare the groups on putative risk factors measured retrospectively. Would this qualify as developmental data? It would if the emphasis were on *when* the disease manifested itself among individuals, and what factors accounted for its occurrence at different times for different people. But most often, the objective is to produce a single estimate of risk (e.g., the relative odds) without reference to any time interval, and in such cases it would not qualify as a developmental study.

Thus, developmental data can be obtained in cross-sectional studies when it is possible to assess and to date individual characteristics or experiences retrospectively. These can produce the same data structure as a longitudinal study, making it possible to describe changes in a response variable over time. This is an attractive strategy for at least two reasons: it is less expensive than a longitudinal study, and loss of sample members occurring at follow-up is not an issue. However, the reliability and validity of information collected retrospectively from respondents are often suspect. The memory of adults for things past is subject to a variety of distortions (Schacter, 1999), and there is some evidence to suggest that the accuracy of retrospective information collected from children and adolescents is particularly suspect (Angold, Erkanli, Costello, & Rutter, 1996; Henry, Moffit, Avshalom, Langley, & Silva, 1994). Thus, the need for longitudinal studies.

Follow-up studies. Follow-up studies involve two occasions of measurement; they are, of course, longitudinal by design and offer the researcher the decided advantage of collecting information prospectively. In many designs the researcher is interested in comparing changes in a response variable for people in different kinds of settings, or for those who have or have not received some intervention. However, the opportunity to examine development as a continuous process is limited because the researcher has information only at the beginning and end of some interval, and no knowledge of fluctuations in the response during the interval. Accordingly, subject response can only be modelled as a linear function of time (a straight line), with a slope equal to the change score between the two occasions. The most significant problem with this approach is that the response variable is measured on both occasions with a

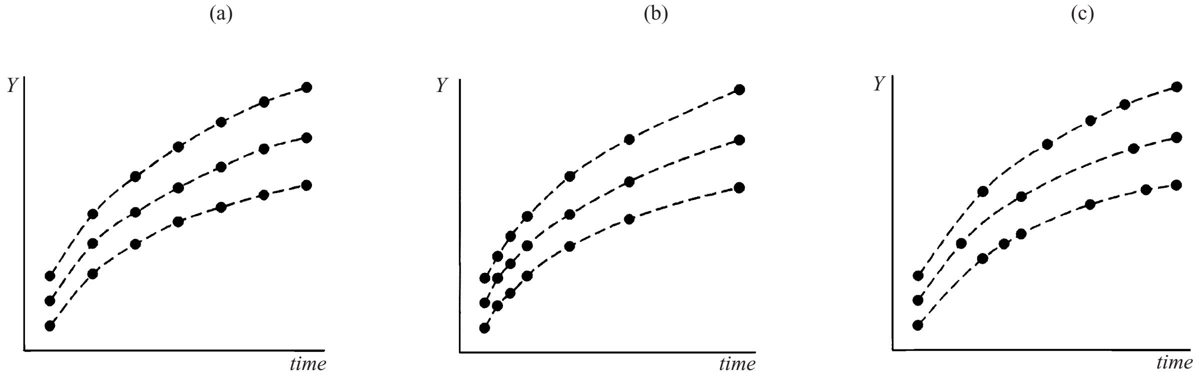

Figure 2. Three types of measurement schedules: (a) fixed, equal spacing; (b) fixed, unequal spacing; (c) variable, subject specific.

certain amount of error. For example, in a typical study, some people receive lower scores than their "true" scores on the first occasion, and higher scores than their "true" scores on the second occasion, making it appear that they had changed considerably. The reverse can hold for other individuals, suggesting that they had not changed very much. Thus, the measurement errors are in a sense compounded, rendering the change scores unreliable (Willett, 1989). This is not a problem when the response variable is measured with a high degree of accuracy; for example, in studies of physical growth where accurate measurements of height and weight can be recorded.

Follow-up studies can also yield reasonable data when researchers are interested in the timing of particular transitions. For example, in an experimental study a researcher might be interested in the remission of depression over a 6-month period among adolescents treated with sertraline vs. cognitive behavioural therapy; or, in an observational study, in adolescents' participation in the labour force subsequent to being classified as completing or not completing high school. In both cases, the researcher might collect data at the beginning of the study and 6 months later. On follow-up, the researcher could ask the respondents to recall onsets of depression or periods of employment, but the same caution regarding the reliability and validity of recall data applies. The advantage in this case, though, is that respondents can be alerted on the first occasion to record their experiences, and then we actually have the longitudinal data we desire, albeit collected by respondents.

Longitudinal studies: Measurement schedule. The study of development as either a transitional or continuous process requires repeated measurement, on at least three occasions, and preferably more. Figure 2 identifies three patterns of data collection for studying continuity and change in individual characteristics: one where the occasions of measurement are fixed and have equal spacing over time (2a); a second where the occasions of measurement are fixed and have unequal spacing (2b); and a third where the occasions of measurement are variable and subject specific (2c).

Studies with occasions of measurement that are fixed have three advantages. Administratively, they are easier to plan and implement because the work and allocation of resources can be streamed evenly over time. They also make it easier to reduce sample attrition, for example, by tracking participants who change residence, and by making regular contact with participants to sustain their interest in the study.

Although fixed data collection intervals are desirable, a more accurate portrayal of development is achieved if data are collected more frequently during periods when the majority of participants are undergoing rapid change. For example, in Fig. 2b measures are taken more frequently during the initial period of rapid growth, such that the nonlinear growth pattern is captured more accurately.

In many studies, however, adherence to a fixed measurement schedule is difficult or even unattainable. For example, fixed schedules are difficult for subjects who are hard to locate or to engage in assessment. In these samples, measurements will be subject specific, and variable from one person to the next. In other studies, it may be impossible to include all subjects in an occasion of measurement because of resource constraints. In this situation, the fixed measurement schedule gets fragmented into different groupings. Finally, most studies experience variable response rates from individual subjects, which leads to missing data for some occasions of measurement. In studies covering long time intervals with several occasions of measurement, it is common for the majority of subjects to be missing data for at least one or two occasions.

Longitudinal studies: Independent vs. clustered sampling. In developmental studies, sampling—the process of identifying and enlisting subjects into a study—provides the means for drawing inferences about broader groups called target populations. With the exception of twin studies, developmental research in the past has focused mostly on the experience of children sampled as *independent* units. For example, in clinic-based studies of child psychopathology, a sample of consecutive referrals, which meet criteria for a targeted disorder, might be asked to participate in the evaluation of an experimental treatment. Clinical investigators designing such intervention trials usually stipulate that enlisted children be unrelated to one another (not from the same family). This is done to insure that responses to treatment are independent from one child to the next. The same strategy has been followed by researchers engaged in general population studies. For example, most of the well-known, contemporary, longitudinal studies of child psychopathology have sampled children as independent units (e.g., see Koot, 1995).

Researchers sampling intact clusters or naturally formed groupings in the past have often done so to facilitate access to individuals. Analytically, these sampling units were seen as nuisance phenomena—reducing the costs of getting subjects at the price of statistical inefficiency (Boyle, 1995). This perception is changing with the recognition that naturally formed groupings are substantively interesting in their own right and contain important information on contextual factors that influence child growth and development. As noted earlier, contextual factors are higher-order or *macro* level variables shared in common by naturally formed groups of individuals. For example, the family defines a naturally formed group of children (siblings) who share in common a variety of structural and process variables expected to differ from one family to the next and to exert variable influences on child growth and development. The neighbourhood defines a naturally formed group of families who share in common a variety of structural and process variables expected to differ from one neighbourhood to the next and to exert varying contextual influences on growth and development, either directly on the child, or indirectly through the family.

There are two important ways in which the study of contextual influences on child growth and development are enhanced by sampling naturally formed groupings. First, assessment data collected from individuals within each naturally formed group can be pooled or aggregated to measure contextual variables hypothesized to influence development. Many contextual variables of interest to researchers such as family functioning, neighbourhood cohesion, or school climate will only be available for study if they can be assessed reliably by aggregating individual assessments. The aggregation of individual assessments generates *macro*-level variables to represent the shared experiences of individuals within groups. The same assessment can be, and often is, used on its own as a lower-level variable to measure individual perceptions (nonshared experiences).

Second, the availability of naturally formed groupings for study makes it possible to estimate the total variation in response that is attributable to contextual influences. Think of a study of child behaviour that includes sibships within families. Having behavioural data on children within families makes it possible to distinguish between-family variation in child behaviour from within-family variation in child behaviour. The estimate of between-family variation in response identifies the extent to which family variables can exert contextual influences on child behaviour. If there is no between-family variation in response, then family-level variables have nothing to account for. The same argument applies to higher-order groupings such as neighbourhoods or schools. Intentionally sampling these units provides the basis for allocating total variation in response across sampled levels so that their overall importance for growth and development can be estimated.

Choice of Analytic Method

Up to this point, we have characterized the study of development as either a continuous or transitional process. It requires longitudinal data for a response variable that is either continuous or discrete, measured on at least three occasions and preferably more. The sample can include subjects sampled as independent units, or from clusters.

Development as a continuous process. Developmental studies that sample children as independent units, collect longitudinal data according to a fixed, equally spaced schedule, and measure *continuous response variables* can be analyzed using traditional statistical methods derived from ANOVA (O'Brien & Kaiser, 1985). Indeed, under certain conditions, trend analysis using ANOVA will yield identical empirical results to the approach recommended in this report: growth curve analysis. These conditions include an identical measurement schedule for all subjects, no missing information, and covariates measured as discrete variables.

A comparison of ANOVA and MANOVA approaches to longitudinal data with growth curve analysis has been provided elsewhere (Francis, Fletcher, Stuebing, Davidson, & Thompson, 1991). Very briefly, the design requirements associated with growth curve analysis are very flexible: predictors of growth can be either discrete or continuous, time dependent or time independent; the number and spacing of time points can vary across subjects; and all subjects with any data can be included in the analysis. In addition, growth curve analysis offers the conceptual advantage that it models growth as a continuous process of within-subject change whereas traditional methods model growth as a series of increments and decrements experienced by groups.

Developmental studies arising from cluster samples or the allocation of naturally formed groupings to interventions pose an additional layer of complexity that goes beyond the scope of traditional statistical methods. As noted earlier, the relational qualities of hierarchical data structures arising from sampling naturally formed groupings pose the same statistical challenges as repeated measurements taken on the same individual. In both instances, responses are correlated. Multilevel or hierarchical linear modelling, developed specifically to account for correlated response variables at multiple levels, provides an efficient and valid approach to the analysis of developmental data for subjects affiliated with naturally formed groupings such as families, neighbourhoods, or schools (Bryk & Raudenbush, 1992; Goldstein, 1995).

In this paper we discuss growth curve analysis for continuous response variables in the context of multilevel models (Bryk & Raudenbush, 1987). We also discuss its extension for hierarchical data, which we refer to as multilevel growth curve analysis. There is room for confusion here: the fact that growth curve analysis for developmental data describing individuals sampled as independent units (i.e., not nested in families, schools, or neighbourhoods, for example) employs a multilevel approach (and is analyzed using the same software), is really beside the point. Its aim is to characterize the growth process and discern which factors affect that process. Its multilevel extension asks the same questions, but allows for data that are nested hierarchically within some context such as a family, school, or neighbourhood.

Development as a transitional process. The most common approach for analyzing development as a transitional process is survival analysis (Greenhouse, Stangl, & Bromberg, 1989; Singer & Willett, 1991, 1993; Willett & Singer, 1993). We refer to the extension of survival analysis to the study of individuals sampled within natural groupings as multilevel survival analysis. Survival analysis was developed originally to study nonreversible or one-way single-state transitions. However, researchers are generating new applications of survival analysis methods to accommodate research questions that focus on transitions to competing states as well as reversible phenomena (Hartmann, Schulgen, Olschewski, & Herzog, 1997; Keiding, 1999; Willett & Singer, 1995).

Although there are numerous examples of studies that employ multilevel models to analyze cross-sectional data, especially studies of schooling (e.g., Raudenbush & Willms, 1991), there are fewer examples of studies that have modelled the development of children sampled from naturally formed groupings. This is expected to change as researchers develop theoretical models to account for contextual influences on child development; as publicly sponsored, longitudinal data sets arising from cluster samples become available to investigators; and as statistical methods, such as multilevel modelling, become more familiar to investigators. The other development expected to stimulate the use of multilevel modelling with longitudinal data is the evaluation of prevention initiatives delivered to intact groupings. Just last year, there were published reports of two school-based trials designed to evaluate the prevention of aggressive, antisocial behaviour using programs delivered by teachers in classrooms (Boyle et al., 1999; Conduct Problems Prevention Group, 1999a, b; Hundert et al., 1999).

The next section describes the methods and data requirements associated with growth curve analysis and its multilevel extensions. It is followed with a section on discrete-time survival analysis and its multilevel extension.

Modelling Development as Continuous Process: Growth Curve Analysis

Growth curve analysis provides a flexible and powerful approach to modelling development as a continuous process. When specified as a general multilevel model, it is demonstrably superior to traditional statistical approaches used by researchers to study change (Francis et al., 1991; Manor & Kark, 1996; Nich & Carroll, 1997; Speer & Greenbaum, 1995). The analytical objective focuses on two activities: (1) developing a statistical equation to summarize the starting point and the trajectory followed by the response variable, which is measured repeatedly on each individual (within-subject growth); and (2) determining the extent to which the starting points and shapes of the response trajectories vary as a function of one or more other measured variables that are used to differentiate individuals (between-subject growth).

Modelling Within-subject Growth

The basic strategy for modelling within-subject growth uses time as the independent variable for predicting response. Thus, the idea underlying growth modelling for an individual is to estimate the person's baseline (starting point) and trajectory (shape of the curve) formed by consecutive assessments taken on the same individual.

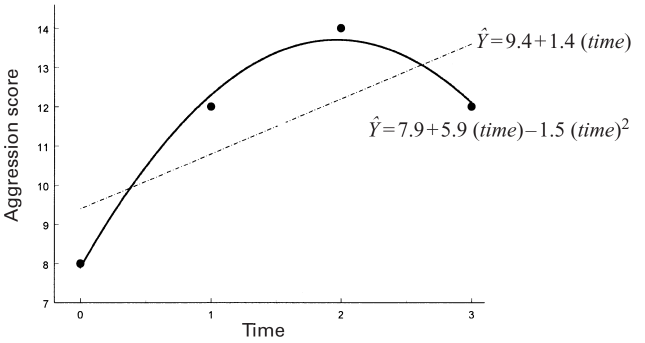

Figure 3. Plot of observed aggression scores versus time, and linear and quadratic growth trajectories.

Figure 3 serves to illustrate some of the basic features of modelling within-individual growth. Imagine that the observed score values arise from measures of aggression repeatedly assessed on a single child. In this case, the child had a score of 8 at the beginning of the study, and scores of 12, 14, and 12 at the end of the first, second, and third weeks respectively. The basic linear growth model posits that scores have a linear relationship with time:

$$Y_t = \pi_0 + \pi_1\,(time)_t + r_t \tag{3}$$

where Y_t is the child's score on occasion t ($t = 1, 2, \ldots T$), π_0 is the estimate of the person's score at baseline, π_1 is the growth trajectory—in this case the slope of the straight line—and r_t is a residual term which indicates the variation of each observation from the hypothesized trajectory on occasion t. This equation is really identical in form to equation (1), except that *time* is the X-variable, and we have changed the Greek letters used to denote the regression parameters. We have done this to differentiate them from β and ε, which are reserved to describe between-person relationships. Similarly, we have used the subscript t to denote the occasion, leaving the subscript i to indicate individuals. In our example, the estimated π_0, π_1 are 9.4 and 1.4 respectively, and the growth trajectory is shown as a dashed line.

We can see that this linear model does not fit the data well, and therefore an alternative model needs to be developed which captures the nonlinear relationship between aggression and time for this child. This can be achieved by adding another parameter to the model so that growth in aggression is expressed as both a linear and a quadratic function of time. The equation corresponding to this model is:

$$Y_t = \pi_0 + \pi_1\,(time)_t + \pi_2\,(time)_t^2 + r_t \tag{4}$$

where π_2 is the coefficient for the quadratic component, and the other parameters are the same as those in equation (3). The estimated values of π_0, π_1, and π_2 in this example are 7.9, 5.9, and -1.5 respectively. The estimated growth trajectory is shown as a heavy solid line.

When modelling individual growth, the number of parameters that can be included is one less than the total number of repeated observations. Adding parameters to a growth model serves the objective of improved model accuracy, but increases the complexity of the model: it should be done when the advantages conferred by improved accuracy outweigh the disadvantages associated with greater complexity. The statistical reliability associated with additional parameters (i.e., statistical significance of the πs) provide a helpful guide for determining whether or not the added complexity is worthwhile. It is not the only consideration, however, especially in studies with very large sample sizes which have the statistical power to confer significance on small effects. In some studies when growth is increasing or decreasing exponentially over time, it may be more efficient to transform the value for time (create an alternate specification) so that response is a single linear function of this transformed value. However, this would not work in Fig. 3 because the subject's response is increasing at first then decreasing midway through the period of observation.

An inspection of Fig. 3 reinforces some earlier comments about growth curve modelling and introduces a new point. First, if the researcher had conducted a follow-up study (i.e., with only two time points), and based findings on data collected at time 0 and time 3, the growth trajectory would simply be a straight line from 8 to 12. This would not only be unreliable because of compounded measurement error, it would be a misleading representation of the true underlying trajectory. Second, growth curve analysis also imposes no restrictions on the measurement schedule. The model for the trajectory in Fig. 3 arose from four waves of data collected at fixed intervals. The same model could result from innumerable combinations of four waves collected at variable times throughout the observation period. Every subject could have a different measurement schedule: it would simply mean that the values for time used to model their individual growth trajectories would vary from one to another and be person specific. Third, growth curve analysis proceeds with the data available on each individual: missing occasions of measurement attenuate the reliability (and validity) of individual growth estimates but do not prevent their computation. Fourth, in situations where individual growth estimates are derived from varying numbers of responses, estimates based on fewer responses are drawn towards group values. The idea is to borrow strength from the data available in an effort to have a more powerful analysis (Kreft & De Leeuw, 1998, p. 14). Accordingly, individual estimates are "shrunken" to the overall solution.

A careful examination of the response trajectories for individual subjects is the point of departure for modelling within-subject growth. This should lead to a correctly specified model of growth for comparing differences among subjects. The advantages of having more occasions of measurement to describe growth within a study observation period have been illustrated in a single example. The selection of a study period will depend on the research question and the hypothesized time course for response. Having four occasions of measurement provides a basis for identifying nonlinear associations between time and response. Researchers often choose equally spaced occasions of measurement, but this is not the optimal strategy when growth is uneven throughout the study. In this situation, it is better to sample observations more frequently during periods of rapid change. Ultimately this decision must be based on both theory and available empirical evidence about the patterns of within-subject change.

Modelling Between-subject Growth

The basic strategy for modelling between-subject growth requires an extension to the approach used to model within-subject growth. Now, there are two models, one that describes within-subject growth and another that describes between-subject growth. Model 1, often called the "within-subject" or "level-1" model can be rewritten from the OLS regression in equation (3) defining within-subject growth for each child as:

$$Y_{it} = \pi_{0i} + \pi_{1i}\,(time)_{it} + r_{it} \qquad (5)$$

where Y_{it} is the score for the *i*th child ($i = 1, 2 \ldots$ n) on occasion t ($t = 1, 2 \ldots \mathrm{T}_i$). This equation is identical to equation (3) except that a subscript *i* has been added to the parameters and the values of *Y* and (*time*). There is another subtle but significant difference from equation (3): the number of occasions, T_i, is subscripted with an *i* to indicate that not all children need to be assessed at the same time, or for the same number of occasions. Thus, model (5) actually specifies *i* equations, one for each subject. Therefore we conceive of estimating a separate trajectory for each child, and thereby having *i* separate estimates of π_0 and *i* separate estimates of π_1.

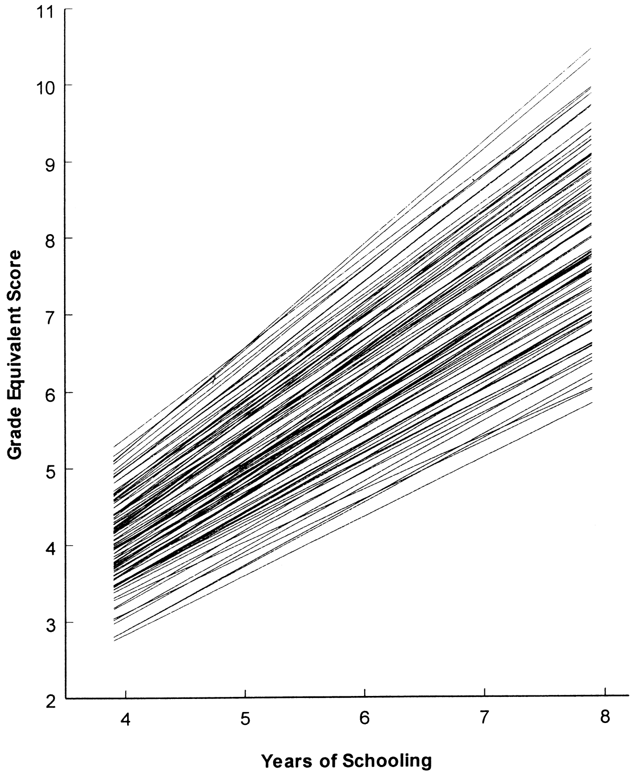

Figure 4. Growth trajectories of children's mathematics scores.

The fundamental idea underlying multilevel modelling is that the parameters from one level become the outcome or response variables at the next level. In this case we can write a between-person equation for the *i* different estimates of π_0, and a between-person equation for the *i* separate estimates of π_i:

$$\pi_{0i} = \beta_{00} + \varepsilon_{0i} \qquad (6)$$

$$\pi_{1i} = \beta_{10} + \varepsilon_{1i} \qquad (7)$$

These equations are virtually identical to our basic OLS regression model—equation (3)—with the "initial status" as the response variable in equation (6), and the "rate of growth" as the response variable in equation (7). The parameters β_{00} and β_{01} represent the average initial status and the average rate of growth respectively. These are important, as they define the overall (or average) growth trajectory. But the growth trajectory for each child deviates from this overall growth trajectory. The residuals ε_{0i} capture the differences between each person's estimated initial status and the average initial status, and the residuals ε_{1i} capture the difference between each person's rate of growth and the average rate of growth.

An example. In a study of school achievement in one school district in British Columbia, the Canadian Test of Basic Skills (CTBS, which is the Canadian version of the Iowa Test of Basic Skills) was administered to students every year from the end of grade 3 through to the end of grade 7 (Willms, 1998; Willms & Jacobsen, 1990). Data were collected with a fixed equally spaced measurement schedule, except that there were missing data because some children were absent during testing, and there were children transferring in and out of the district over the course of the study. Figure 4 displays the growth trajectories for a random sample of 100 of the students from the district. The vertical axis is a measure of students' mathematics achievement, expressed as grade equivalent scores. The horizontal axis is a measure of the number of years of schooling the child had received since kindergarten, irrespective of whether they had progressed normally or had repeated a grade. Thus, a child who was progressing at a rate similar to the national average would have a grade equivalent score of 4.0 after 4 years of schooling, and would increase his or her score at a rate of 1 grade equivalent per year. The average "initial status" (i.e., β_{00}) for this district at the beginning of grade 4 was 4.15, which is about $1\frac{1}{2}$ months of schooling above Canadian norms. The average rate of growth (i.e., β_{01}) over the 4-year period was 0.995 grade equivalents per year, which is very close to the expected growth rate of 1.0 grade equivalents per school year.

The figure shows clearly that the initial status of these 100 children differ from one another, as do their rates of growth. The residuals ε_{0i} and ε_{1i} are expected to have a mean of zero and variances σ_0^2 and σ_1^2. Estimates of σ_0^2 and σ_1^2 are centrally important because they quantify the amount of variation between subjects in their initial status and rates of growth. Between-subject variation in rates of growth is a prerequisite for studying factors hypothesized to influence growth. In this example, if σ_1^2 had been zero, the lines would be parallel, and we could not model variation in children's growth.

The residuals $\varepsilon_{0\mathrm{i}}$ and $\varepsilon_{1\mathrm{i}}$ also covary, and have a covariance of σ_{01}, or in less technical terms, there can be a correlation between children's initial status and their rate of growth. In this example, the correlation is positive, and we observe a "fan-spread" pattern—in essence the proverbial "rich get richer, and poor get poorer" phenomena. In some instances, the correlation could be negative—a "fan-close" pattern—or there could be no correlation between growth and initial status (see Bryk, 1980).

If you find all the Greek letters and subscripts getting in the way, push them aside for a moment and remember the following. The method of growth curve analysis is simply regression with a few twists. The statistical equations in (5), (6), and (7) are regression equations: the first one examines within-subject responses over time, baseline and trajectory; the second two examine between-subject responses over time for the estimated baselines and trajectories from the first equation. The error or residual terms are subdivided into within-subject terms and between-subject terms. The most important difference between "regular" regression and regression formulated as growth curve analysis is the existence of separate error terms at level 1 and level 2.

Expanding Model Capacity

Growth curve analysis has the same model-building flexibility as regular regression. Additional variables can be added to the equations to serve a variety of objectives: (1) to explain variation in between-subject growth; (2) to control for confounding variables that may distort either within- or between-subject comparisons of primary interest; and (3) to evaluate statistical interactions that might signal moderating effects for hypothesized variables. One of the exciting features of growth curve modelling, however, is its ability to explicitly include time-invariant and time-variant covariates.

Adding time-invariant covariates. Following the example in Fig. 4, suppose one wanted to examine the effects of the child's sex on growth in mathematics achievement. This is done simply by adding a "dummy" variable denoting sex (e.g., 1 for females; 0 for males) to equations (6) and (7):

$$\pi_{0i} = \beta_{00} + \beta_{01}\,(female)_i + \varepsilon_{0i} \qquad (8)$$

$$\pi_{1i} = \beta_{10} + \beta_{11}\,(female)_i + \varepsilon_{1i} \qquad (9)$$

The estimate of β_{01} for equation (8) indicates the average difference between girls and boys in their initial status. (We find it convenient to call the variable "female" rather than "sex", as it then indicates that the variable was coded 1 for girls and 0 for boys, rather than the other way around.) Similarly, the estimate of β_{11} for equation (9) indicates the average difference between girls and boys in their rate of growth. As in OLS regression, when covariates such as this are added to a growth curve model, their estimates are adjusted for all other variables in the model.

Note that β_{00} and β_{01} in equations (8) and (9) now represent the expected initial status and rates of growth for *boys*, rather than the average child as in equations (6) and (7). In multilevel modelling, a convenient way to think about an intercept is that it is the expected score for an hypothetical person who has a score of zero on all of the covariates. In this case, a person with zero on the covariates would be a boy. Covariates such as age of the child or the educational level of a child's parents are often "centred" by subtracting the score or value for each individual from the overall mean score or value. The intercept of the between-child equation for initial status then indicates the expected score for a child who was of average age, and whose parents had an average level of education, rather than a child aged zero with parents who had no education. The same applies for the model for rates of growth. Note that centring only "shifts" the value of the intercept to lend it meaning; it does not affect the magnitude or shape of the growth trajectories. The coefficients for continuous variables are interpreted in the same way as in OLS regression: they indicate *the expected change in* Y *for a one-unit change in the covariate, given that all other covariates in the model are held constant.* In most cases, the coefficients are more easily interpreted if the covariates are scaled in a meaningful metric (e.g., years of education, family income in $1000 units). Thus, in most instances we centre covariates (by subtracting their mean), but do not standardize them (by dividing the centred variable by its standard deviation).

Adding time-variant covariates. Growth curve modelling is not restricted to time-invariant covariates, and this is an important asset. For example, if the investigators of the study described above had collected data on whether a child had changed residences, and if so, *when* they had moved, a time-dependent variable assessing whether or not a child moved during each interval could be included. A time-varying covariate would be added to model (5) above:

$$Y_{it} = \pi_{0i} + \pi_{1i}\,(time)_{it} + \pi_{2i}\,(moved)_{it} + r_{it} \qquad (10)$$

where $\pi_{2\mathrm{i}}$ represents the effect of moving for the *i*th child. If this could be shown in Fig. 4, we would see straight lines for most children, but for those who had moved during the course of the study, there would be a shift, either up or down, in their growth trajectory. The multilevel model would include a second-level equation, similar to equation (6), which indicated the average effect of changing residence.

When we began the discussion of growth curve modelling we indicated that the within-person model does not have to be linear. We showed with our first example (Fig. 3) that the within-person model might include a quadratic term. The quadratic term is handled in the multilevel context in the same way as any time-variant covariate: there is an extra term in the within-person model, and an equation at the second level like equation (6) that expresses the within-person coefficients as an average plus residuals. The only difference is that the parameter indicates the extent of nonlinearity.

The introduction of a covariate to the within-person model takes up one extra "degree of freedom", just as adding a quadratic term does. Thus, if researchers wish to include several time-variant covariates, they need several measurement occasions. Growth curve analyses provides enormous opportunities for examining developmental influences. The limiting considerations for selecting independent variables arise from the study itself, including the strength of its design, the adequacy of measurement, and statistical power afforded by the study sample.

Assessing intervention effects. If some of the children in Fig. 4 had been exposed to a new mathematics program, and others had received the traditional program during the course of the study, one could assess its "intervention effect" by including a dummy variable (1 for those "exposed", 0 for "controls") to the between-person model, in exactly the same way that "female" was included in equations (8) and (9):

$$\pi_{0i} = \beta_{00} + \beta_{01}\,(female)_i + \beta_{02}\,(newmath)_i + \varepsilon_{0i} \quad (11)$$

$$\pi_{1i} = \beta_{10} + \beta_{11}\,(female)_i + \beta_{12}\,(newmath)_i + \varepsilon_{1i} \quad (12)$$

The estimate of β_{12} is of primary interest, as it indicates the effect on children's rates of growth associated with receipt of the intervention. However, β_{02} is important also, as it indicates whether children who received the new program differed from others at the onset of the study. This type of design provides a much stronger control for "selection bias" than cross-sectional or follow-up studies, but it does not rule it out.

As with OLS regression models, one can also assess interactions among variables. In the example above, the researcher can add an interaction term, *female-by-newmath*, to discern whether the program differed in its effects for males and females. One can also include more than one treatment variable in the model, or even assess interactions among two competing treatments.

This model provides a powerful means of assessing treatment effects. An even more powerful model can be achieved if the treatment is introduced at different times for different sample members. The treatment variable is then a time-variant covariate, and is added to the within-person model. The equation is identical to equation (10) except that one has a variable denoting whether and when a person received the intervention, instead of whether and when they changed residences. This model is more powerful in that each person essentially serves as his or her own control group. If one can detect a discernible shift up or down in the growth trajectories, following the period people received the intervention, it would be difficult to argue that these effects arose from selection bias or period effects. We recommend this type of design when random assignment is impossible. It is also a useful design for experiments where it is politically unacceptable to withhold treatment, because with this design, everyone can eventually receive treatment.

Growth Curve Modelling with Hierarchical Data

It is not too difficult to imagine statistical dependence or correlations among identical measures repeated over time on the same subject. It is more challenging to consider statistical dependence among measures taken on individuals belonging to naturally formed clusters; think about children within families, and patterns of similarity between identical twins, fraternal twins, natural siblings, siblings created through reconstituted families, and non-siblings. In this situation, one can expect a hierarchy of correlated responses that flow from similarity in genetic background and family environment. In neighbourhoods and schools, one can also expect that shared experiences will lead to correlated responses, albeit lower in magnitude than either repeated measures on the same individual or measures taken among siblings in the same families.

Using the platform of multilevel or hierarchical linear modelling, it is relatively straightforward to extend growth curve analysis to accommodate correlated responses that occur when individual subjects are sampled as part of naturally formed clusters or affiliations. This requires a three-level model, with the clusters at the highest level (see Raudenbush & Bryk, 1988). The first two levels are identical to those described by equations (5), (6), and (7). They are repeated here with a subscript that denotes membership in a particular cluster j ($_{\mathrm{j}} = 1, 2 \ldots n$):

$$Y_{ijt} = \pi_{0ij} + \pi_{1ij}\,(time)_{ijt} + r_{ijt} \quad (13)$$

$$\pi_{0ij} = \beta_{00j} + \varepsilon_{0ij} \quad (14)$$

$$\pi_{1ij} = \beta_{10j} + \varepsilon_{1ij} \quad (15)$$

Notice that our average initial status, β_{00}, and average rate of growth, β_{01}, now have a subscript j. This is because we now have j sets of these parameters, one for each cluster. We want to know whether these vary among clusters, and if so, whether we can explain this variation with cluster-level variables. The β_{00}s and β_{01}s become the dependent variables in regression models at the third level:

$$\beta_{00j} = \phi_{000} + U_{00j} \text{ (between clusters in average initial status)} \quad (16)$$

$$\beta_{10j} = \phi_{100} + U_{10j} \text{ (between clusters in average rates of growth)} \quad (17)$$

This three-level model may be clearer with a concrete example. The growth data displayed in Fig. 4 are actually a subsample of data collected for over 1500 children in 31 elementary schools from 1 school district. Figure 5 displays the average growth trajectories for the children within each school—these are the β_{00j} and β_{10j} ($j = 1, 2, \ldots 31$) of equations (14) and (15). The results show that most schools have an average growth rate close to 1.0, which is consistent with Canadian norms. However, students in the district's three worst-performing schools were on average growing at a rate of about 0.85 grade equivalents per year, whereas the students in the three best performing schools were growing at a rate of about 1.11 grade equivalents per year. The β_{00j} and β_{10j} become the dependent variables in equations (16) and (17), and these models can be extended to include covariates

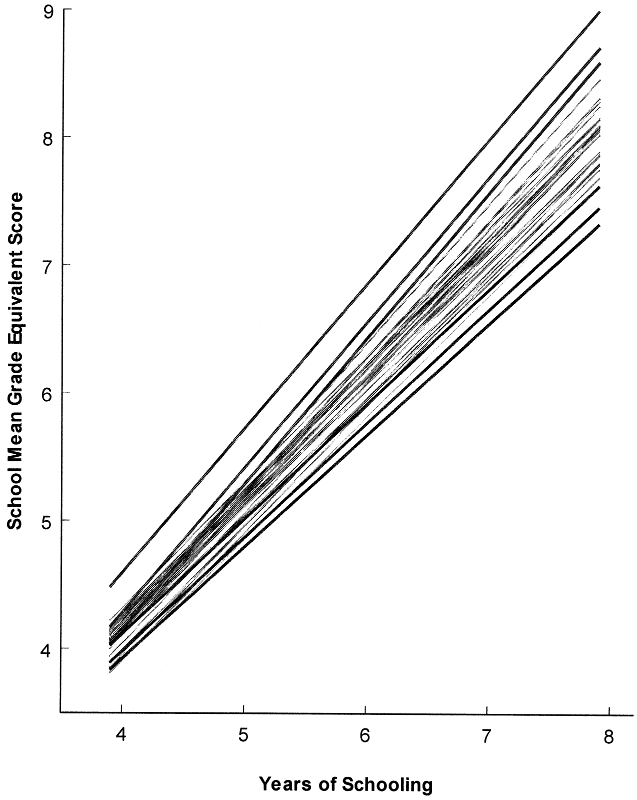

Figure 5. Average within-school growth trajectories for 31 elementary schools.

describing school resources and school policy and practice (e.g., pupil-teacher ratio, retention policies, ability grouping). The analysis of growth in this manner can provide a very accurate and powerful means of assessing the "effect" or "added value" associated with attendance or service from a particular institution. Usually, the between-person models [(14) and (15)] include covariates describing the person's sex and socioeconomic status, as a means to control for differences associated with direct and indirect selection into particular institutions. The between-cluster models [(16) and (17)] include covariates describing policy and practice, or some intervention that was introduced into a few of the institutions (see Raudenbush & Willms, 1995).

The Greek letters and subscripts are provided for the mathematically inclined. Don't let them obscure the message. The message is that growth curve analysis focused on between-subject variation in growth among "independent" samples of children can be extended to examine between-group variation in growth of naturally formed clusters. The extension from a two-level model to a three-level model is basically the same as the extension from a one-level to a two-level model, covered earlier in some detail. With sufficient data at multiple levels, this extension can push upward, hence the name multilevel or hierarchical linear models.

Modelling Development as Transitional Process: Discrete-time Survival Analysis

The previous section examined development as continuous process. Attention focused on quantitative response variables such as school achievement, intelligence, mood, or aggressive behaviour. This section examines development as transitional process. There are many changes and events during infancy, childhood, and adolescence of interest to a variety of developmental researchers. For example, the timing of developmental milestones such as a child's first use of language or walking have important implications for the subsequent acquisition of verbal and motor skills. The study of developmental psychopathology focuses on the acquisition of knowledge about the onset and course of childhood psychiatric disorder. Timing the occurrence of disorder onset is an important step in the study of etiology; timing the occurrence of recovery is an important step in the evaluation of treatment. Life course researchers pay special attention to important transitions such as entry into the work force and the initiation of family life, because the occurrence and timing of these events have important implications for individuals' life quality in the years ahead. All of these researchers are interested in addressing the same three questions: (1) Over the course of time, which individuals experience these transitions? (2) Among those who experience them, when do they occur? and (3) What factors affect whether and when people experience these transitions?

The methods used to investigate status change and event occurrence go by a variety of names, including survival analysis, event history analysis, and hazards modelling. These methods were developed by researchers examining life expectancy in human populations (Cox, 1972; Cox & Oakes, 1984; Kalbfleisch & Prentice, 1980; Miller, 1981). They were concerned mostly with estimating the cumulative probability of survival from birth, displayed against time as a monotonically decreasing curve or step function (vertical drops after a specific interval). The use of data describing various characteristics of individuals, such as sex, residence, and year of birth, to create mutually exclusive and collectively exhaustive subgroups made it possible to assess the association between independent or predictor variables and survival over time. This was done by comparing the survival among the groups, in particular, the precipitousness of the curves or steps which applied to them. With death serving as the primary end point, the concepts attached to these methods (and still used) took an ominous tone. These include the idea of *risk period*, which refers to an interval during which an individual is at risk for "death", and the idea of *hazard probability*, which refers to the chance or likelihood of "death" occurring during a particular interval among those alive at the beginning of the interval. Given the reasonableness of certain assumptions, it is possible to substitute for death, any transition, status change, or event occurrence, and use these methods to study occurrence and timing.

Plotting Subject Transitions

The point of departure for modelling development as a transitional process is an examination of subject response over time, and this can be illustrated by example. Suppose that a clinician wants to examine recovery from depression among adolescents diagnosed and treated in an outpatient department with either anti-depressive medication (MED) or cognitive behavioural therapy (CBT). The clinician decides to screen consecutive referrals for depression, enrol adolescents who are eligible for treatment into MED or CBT, and monitor their progress on a monthly basis between January and July, 2001. Suppose that the clinician enrols 20 adolescents into the study and draws a line for each one to indicate the time when they

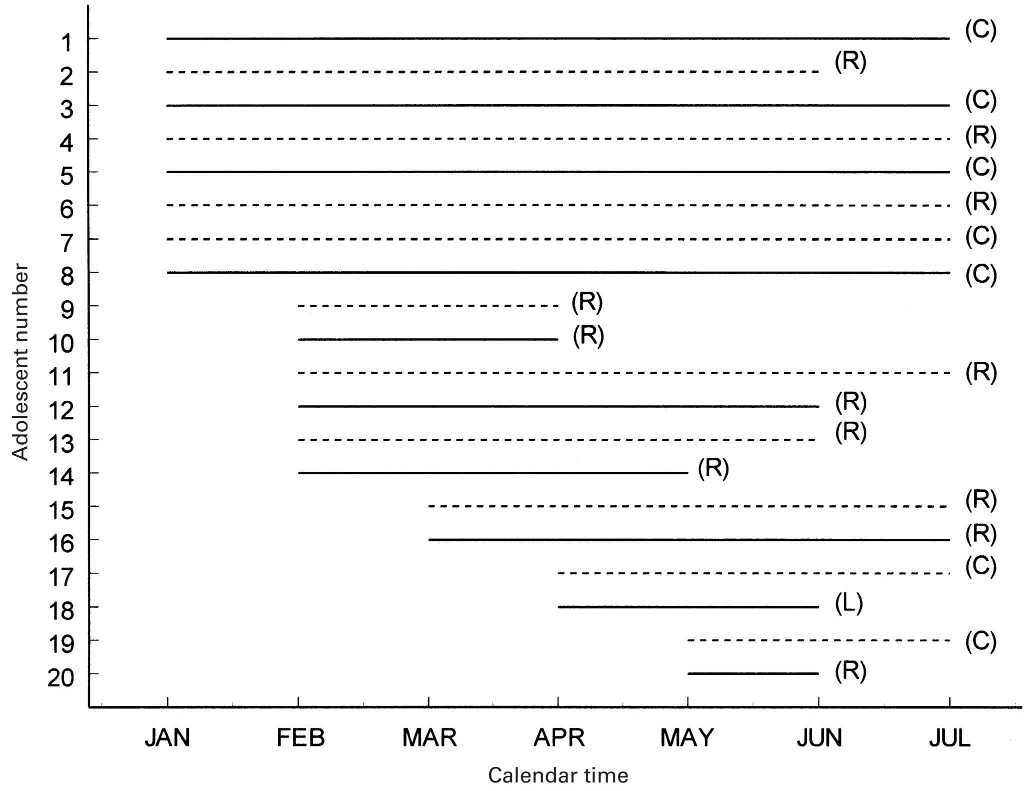

Figure 6. Hypothetical experiences of 20 adolescents allocated to treatment with anti-depression medication (solid line) versus cognitive behaviour therapy (dashed line). Study period January to July. Responses are recovered (R), lost to follow-up (L), and still not recovered by the end of the study period, which is called right censored (C).

Table 1
Data for Estimating Hazard, Survival, and Cumulative Survival for Clinic Example

Risk interval	No. of adolescents at risk	Recovered	Lost/censored	Hazard	Log odds hazard	Survival	Cumulative survival
0–1	20	1	0	.050	−2.94	.950	.950
1–2	19	2	2	.105	−2.14	.895	.850
2–3	15	2	0	.133	−1.87	.867	.737
3–4	13	3	1	.231	−1.20	.769	.567
4–5	9	2	0	.222	−1.25	.778	.441
5–6	7	2	5	.286	−0.91	.714	.315

entered the study and the time when the study period ended for them. The study period would end for an individual when: (1) the adolescent recovered from depression—the transition of interest; (2) the adolescent was lost to follow-up before the overall study formally ended in July—a study withdrawal without the occurrence of a transition; or (3) the adolescent had not recovered before the overall study formally ended in July—in essence also a study withdrawal without the occurrence of a transition, which is called right censoring.

Subject responses over time in this study appear as line drawings in Fig. 6. They illustrate some important concepts. For example, the "study period" is defined in calendar time as the beginning and end of the overall study: it is shown at the bottom of the figure. The "risk period" is defined separately for each individual as the amount of person time contributed to the study until he or she experiences the transition or is withdrawn: it is shown by the lines inside the figure which characterize the length of time from study entry (beginning of the lines) until the study ends for that adolescent. Although adolescents have entered the study at different calendar times, the study clock defining their period of risk starts at a common predefined point: when they enter into treatment. The letter at the end of the line gives the status or disposition of each adolescent at the end of his or her risk period (i.e., when the study ends for him or her). The letter R stands for recovery, the letter L for lost to follow-up, and the letter C for right censored. Lost to follow-up are subjects excluded from the study before its completion: their status is unknown beyond the point of exclusion. Right censored are subjects who are known not to have recovered at the completion of the study: their status is unknown beyond the study period. The data contributions from subjects lost to follow-up or right censored end before the study outcome can be observed. The removal of these individuals from the study has to be accommodated properly in the analysis—a feature that distinguishes survival analysis from other regression-based approaches.

Estimating Hazard Probabilities from Observed Data

Table 1 displays the risk periods applicable to the 20 children in Fig. 6. Instead of calendar time, their experiences are shown from their common starting points (initiation of treatment), and are displayed in monthly intervals. The intervals represent discrete risk periods, not exact dates for recording subject transitions or responses, and provide the means to define and estimate a particularly important statistic: the "hazard probability". *The hazard probability is the chance or likelihood of a study outcome occurring during a particular risk period.* For example, in the first study interval 0–1 months, 20 adolescents are "at risk" for recovery at the beginning of the interval; 1 adolescent experiences recovery during the interval; and 1 adolescent is lost to follow-up during the interval. Accordingly, the hazard probability associated with this interval is 1/20 or .05. This means that adolescents with depression at the beginning of the first interval have a .05 or 5% chance of recovery. (It is customary to assume that subject withdrawals, when they occur, take place at the midpoint of an interval so that their person time contributed to the study is divided in half. To simplify the example, we assume that subject withdrawals occur at the end of the interval). In the third interval 2–3 months, 15 adolescents are "at risk" for recovery at the beginning of the interval; 2 adolescents experience recovery during the interval; and none are lost or censored. The hazard probability associated with this interval is 2/15 or .133.

The hazard probability serves many important objectives. It identifies the peak periods of transition or recovery: for this example, these occur during the fourth, fifth, and sixth intervals. The hazard probability properly accounts for subjects who are removed from the analysis before experiencing the study outcome due to loss to follow-up or censoring: it does this by including their risk period or person time of observation in the denominator of each applicable risk interval. Hazard probability is used to estimate other parameters of interest to researchers. For example, its complement (1 minus the hazard probability) defines the probability of survival during a risk interval. This is given in column 7 of Table 1. The probability of survival during each interval is used to define the cumulative probability of survival. This is called the *survival function*, from which survival curves are plotted. *The cumulative survival is the chance or likelihood that a randomly selected individual from a defined population or sample will "survive" (i.e., not experience the transition) up to a particular time point.* This probability is conditional on "surviving" earlier periods. In Table 1, the cumulative probability of surviving the second risk period (month 1 to 2) is .85, which is derived from .95 times .895. It is the joint probability of surviving the first two risk intervals. The hazard probability is used also to define the *hazard function*, which is a plot of hazard probabilities in order from the first to the last risk interval. Statistically modelling hazard functions and comparing profiles associated with independent or predictor variables are specific objectives of these types of analyses.

The statistical models used to summarize the occurrence and timing of transitions are based on several assumptions that should be reasonable if the analyses are to be trusted. First, subjects who are removed from analysis because of loss to follow-up should be few in number and lost at random. In the study of child development as either a continuous or transitional process, the assumption that subject losses occur at random can never be taken for granted: it must be evaluated. The importance of doing this evaluation is related directly to the magnitude of subject losses. This evaluation can be achieved by modelling subject status (i.e., retained or lost to follow-up) based on information known or believed to be related to subject response or outcome. Second, the time-dependent profiles arising from the pattern of hazard probabilities are bound by the cumulative period of risk in which subjects are observed. It seems reasonable to assume, for example, that adolescents lost to follow-up or censored in Table 1 will experience recovery at some time in the future. However, the occurrence and timing of recovery beyond 6 months cannot be known without additional follow-up of those censored from the analysis. Third, it is assumed that secular trends or phenomena that occur in calendar time have no influence on hazard probabilities. This could occur if a change in criteria for treatment eligibility occurred in March and shifted the recovery profile for one treatment group but not the other. Fourth, the study outcome must be clear and well defined. Most often it will take the form of a binary variable indicating a status change or an event occurrence. It can, however, be defined as a transition into competing states such as the transition from school into full-time employment, part-time unemployment, or unemployment. Finally, there should be a clear and well-defined starting point that is generally applicable to each subject and relevant for answering the particular research question. The clinician in the present example wished to study adolescent recovery from depression after treatment initiation. If someone wished to study the course of adolescent depression from symptom onset, it would be necessary to identify the first onset of depression in a group of adolescents, and monitor their response over time.

Graphing Hazard Probabilities from Observed Data to Reveal a Profile

The hazard probabilities and cumulative survival arising from the clinical data in Table 1 are graphed by treatment group (i.e., MED, solid lines; and CBT, dashed lines) in Fig. 7. The profile of hazard probabilities for the MED group (solid line) is flat, falling off to .0 in the interval 5 to 6. In contrast, the profile of hazard probabilities for the CBT group (dashed line) is relatively flat during intervals 1 to 2 and 2 to 3, and then rises quickly during intervals 3 to 4, 4 to 5, and 5 to 6. The cumulative survival function in Fig. 7 tells a similar story, as it must because it is derived from the hazard probabilities, but draws attention to slightly different points. For example, it clarifies differences in time-course to recovery for the two groups. In the MED group, recovery is steeper during the first few intervals, then tapers off. In the CBT group, recovery catches up and then surpasses the MED group during interval 3. By the end of the study period, the cumulative survival function shows that the proportion of adolescents continuing with depression was higher in the MED group (.57) than in the CBT group (.11).

These graphs are very informative about the recovery experiences of the two samples and provide the starting point for developing statistical models for comparing hazard profiles. Statistical models are needed for two

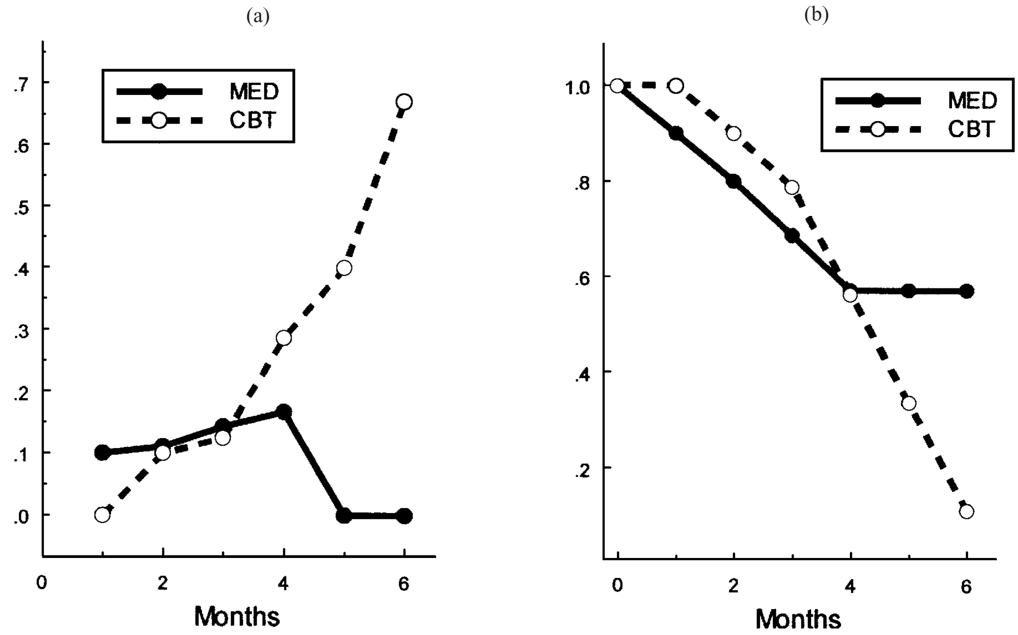

Figure 7. (a) Hazard probabilities and (b) cumulative survival for clinic treatment example.

reasons: (1) to facilitate formal hypothesis testing and confidence interval estimation; and (2) to summarize complex patterns of association that would be unwieldy in graphical form. The latter would come to life for a researcher attempting to evaluate statistical interactions among multiple covariates in a large study.

Statistically Modelling the Baseline Hazard Function

Discrete-time survival analysis in child development research focuses on the same two activities that were discussed under growth curve analysis: (1) developing a statistical equation to summarize the relationship between time and the probability of observing a study outcome (defining the baseline hazard function or profile of risk probabilities); and (2) determining the extent to which the hazard function "shifts" as a function of one or more other variables used to differentiate subjects. Unlike multilevel models, fitting discrete-time hazard models requires no special statistical software: standard logistic regression software can be used. However, the data must be arranged in an appropriate "person-period" format (e.g., see Willett & Singer, 1993).

The statistical approaches used to model hazard probabilities in survival analysis are based on regression methods in which the dependent variable is a hazard function or profile of hazard probabilities and the covariates are subject characteristics or experiences. In our example, the treatment (MED vs. CBT) is the covariate of interest. The most general approach to modelling the hazard function is to construct a statistical equation that includes each and every hazard probability. This is given by:

$$\text{logit}_e\, h_j(t_i) = \beta_0(t) \qquad (18)$$

Referring to the hypothetical study of adolescent depression, $\text{logit}_e\ \text{h}_j(t_i)$ represents the log-odds of recovery for adolescent j during the risk interval i. These are estimated by expressing the hazard probability, i.e., h, as the odds of recovery, i.e., $h/(1-h)$, and then taking the natural logarithm of this odds, i.e., $\ln(h/(1-h))$. In the clinic example, the probability of recovery during the 0–1 month interval was given as .05 (Table 1). The odds of recovering during this interval is $.05/(1-.05)$ or .053, and the log-odds of recovering during this interval is -2.94. Note that unlike the hazard function, which is bounded between 0 and 1 and therefore violates one of the assumptions of OLS regression, the transformed values —the log-odds—have no lower bound. They are shown in the sixth column of Table 1.

Modelling Between-subject Shifts in the Hazard Profile

The basic strategy for modelling between-subject shifts in the hazard profile is to add a predictor or independent variable to equation (18) in the same way that one would in linear regression. The effects of doing this are illustrated using the data from Fig. 6. The modelled hazard profiles for adolescents treated with MED (solid line) versus those treated with CBT (dashed line) are shown in panel (a) of Fig. 8. The equation depicting this association is given by:

$$\text{logit}_e\, h_j(t_i) = \beta_0(t) + \beta_1\ (\textit{intervention})_j \qquad (19)$$

where β_1 represents a shift in the logit hazard profile (either positive or negative) associated with treatment. Note that β_0 is not a constant, as in OLS regression, but a function of time. Thus, in our example, there are six different β_0s, one for each interval. One can take the results from this model, and transform the values back to hazard probabilities, which can be plotted for each interval as was done for Fig. 7. When this model is fit to the data for our example, with the intervention variable coded 1 for MED and 0 for CBT, the estimate of β_1 is -0.77. The resulting hazard profile (after transforming back the logit hazard profile) is displayed in Fig. 8a.

The hazard profiles derived from our statistical model (equation 19), which are displayed in Fig. 8a, differ considerably from the hazard profiles that were plotted directly from the data, which are displayed in Fig. 7a. This is because the statistical model constrained the shape of the logit hazard functions for both groups to be the same at each interval, separated only by β_1. This is called the proportionality assumption. The hazard profiles

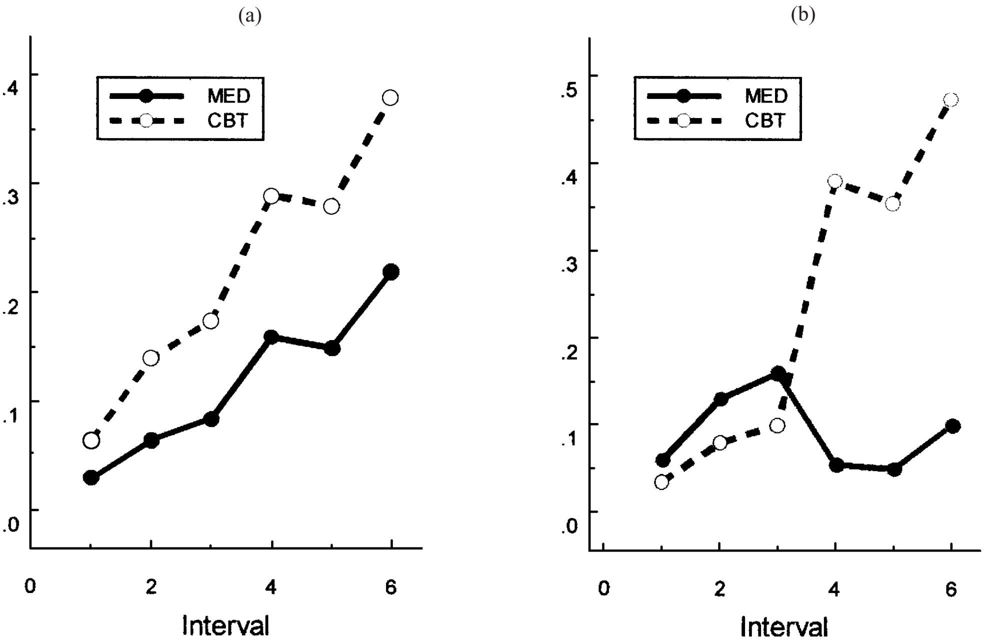

Figure 8. Hazard probabilities for clinic treatment example: (a) displays modelled estimates for treatment; (b) displays modelled estimates for treatment by time interaction.

plotted from the data, however, suggest that the two interventions diverge in their hazard over the course of the study. Whether this difference is statistically significant can be tested empirically; that is, we wish to test the viability of the proportionality assumption. This is accomplished by adding an intervention-by-time interaction term to equation 19:

$$\text{logit}_e\, h_j(t_i) = \beta_0(t) + \beta_1\,(intervention)_j + \beta_2\,(intervention * time)_j \qquad (20)$$

We fit this model to our data, but to simplify the model, *time* was coded 0 for intervals 1, 2, and 3; and 1 for intervals 4, 5, and 6. The analysis yielded estimates of $\beta_1 = 0.56$ and $\beta_2 = -2.81$, with the estimate for β_2 marginally significant at $p = .07$. Thus we reject the proportionality assumption, and conclude that the recovery rates for these two interventions differ over time. Transforming the logit hazard probabilities back to hazard probabilities yields the hazard profiles displayed in Fig. 8 b, which are a considerably better fit to the data than those derived from equation (19).

Expanding Model Capacity

The discrete-time survival models developed in this section can be expanded in ways very similar to those discussed for individual growth curve analysis. The discussion here will be very brief to avoid excessive repetition. The flexibility of individual growth curve analysis and survival analysis arise from their common grounding in regression.

Adding covariates. Covariates in discrete-time survival analysis serve the same function as they do in individual growth curve analysis: (1) to explain additional variation in between-subject risk profiles; (2) to control for confounding variables that may distort between-subject comparisons of primary interest; and (3) to evaluate statistical interactions that might signal moderating effects for particular variables. Covariates can be measured as quantitative or discrete variables; they can be either fixed (time independent) or varying (time dependent); and they can be defined as either *exogenous* or *endogenous*. Statistical interactions can be created to examine the moderating effects of one variable on another. The reader should remember as well that "time" itself is the most important "covariate" in the study of developmental processes.

Specifying alternative effects of time. Whether focused on continuous processes or transitions, correctly specifying the functional relationship between time and response is one of the most important steps in developmental inquiry. In growth curve analysis, the discussion of this issue centred on evaluating more complex specifications of the relationship between time and response, moving from linear to nonlinear representations. In moving from simpler to more complex specifications of this relationship, the objective was to evaluate the potential value of increasing model accuracy against the cost of increasing model complexity. In discrete-time survival analysis, the discussion of this issue moves in the opposite direction. Because the relationship between time and response is specified completely in discrete-time survival analysis, the discussion must centre on evaluating simpler specifications: the objective now is to evaluate the potential value of increasing model simplicity against the cost of decreasing model accuracy.

When we fit equation (18) to our data we treated time as discrete intervals, and estimated the baseline hazard model using a set of six dummy variables, with one dummy for each interval:

$$\text{logit}_e\, h_j(t_i) = \beta_0(t) = \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \alpha_4 D_4 + \alpha_5 D_5 + \alpha_6 D_6 \qquad (21)$$

Thus, our model yielded separate estimates for each interval. Although the general form of this model provides a completely accurate representation of sample risk, it may be unnecessary to use so many parameters. Instead of using six parameters to describe the probability of recovery in the clinic sample, it is possible to use only two: the baseline logit hazard profile would be written as:

$$\text{logit}_e\, \mathrm{h_j(t_i)} = \beta_0 + \beta_1(\mathrm{t}) \qquad (22)$$

In this model, the six parameters used to define the relationship between time and risk in equation (21) have been replaced by two parameters: a baseline risk or starting value and the risk expressed as a function of time measured continuously. $\text{logit}_e\, h_j(t_i)$ still represents the log odds of recovery for adolescent j during the risk interval i. However, it is now predicted by an intercept, β_0, which is an estimate of the hazard probability during the first risk interval and a linear increment, β_1, added for each successive risk interval. Fitting this model to the clinic example yields values of $\beta_0 = -2.94$ and $\beta_1 = -0.44$. Both of these estimates are significant at $p < .01$. As before, with equation (20), this model can be extended to include a main effect for intervention status, and an intervention-by-time interaction term.

Choosing among statistical models to characterise the baseline hazard profile usually involves a trade-off between model simplicity achieved using fewer parameters and model accuracy (goodness of fit between the observed and model predicted values) achieved by using more parameters. When goodness of fit is unaffected by a reduction in the number of parameters, simpler models are preferred to more complex ones. When simpler models degrade goodness of fit, then an evaluation of the trade-off must ensue. Willett and Singer (1993; Singer & Willett, 1993) discuss this model evaluation in more detail. The identification of an appropriate model to express the relationship between time and response can be helped by both theory and previous empirical work but will be constrained ultimately by the amount of data available to an investigator.

Survival Analysis with Hierarchical Data

The model used to construct a hazard profile for an independent sample of individuals can be expanded to include more complex data structures that arise from sampling naturally formed groups or clusters where responses are expected to be correlated. In this instance, the "person-period" format is repeated for each group, such as neighbourhoods, schools, or communities. Just as we were interested in whether the hazard profiles differed for the two groups in the above example, a multilevel survival analysis is interested in whether the results vary among several groups.

Ma and Willms (1999) provide an example pertaining to the likelihood of students dropping out of advanced mathematics over the course of their secondary school career. The data described students nested within schools, and the model covered five periods, grades 8 through 12. The *within-school* survival model included covariates for sex, socioeconomic status, prior mathematics achievement, and prior attitudes towards mathematics; the between-school model included a number of variables describing the social context and academic climate of the school. They found that there were two critical transitions when a large number of students ceased taking mathematics. The first was between grades 8 to 9, and for that transition, prior achievement was the most important predictor. The second transition was from grades 11 to 12, and during that transition, students' attitudes played a more important role. The model also includes some sex-by-grade interaction terms, which revealed that the hazard rate was especially large for females in the transition from grades 11 to 12, and that it was attitudes towards mathematics, not ability, which affected their nonparticipation. The findings also revealed significant differences among schools in their hazard rates, especially for the transitions during the last 2 years of secondary school. This variation was evident even after controlling for students' characteristics and family background. However, their sample lacked a sufficient number of schools to adequately determine why some schools seemed to be able to maintain high participation rates through to the end of secondary school.

Statistical Software

There is a variety of statistical software available to analyze hierarchical data with continuous response variables. A detailed comparison of earlier software based on random regression models was provided by Kreft et al. (1994). These software packages have been updated to model discrete response variables, and a comparison of these packages—MLn, MLwinN, SAS Proc Mixed (Glimmix Macro), HLM, and VARCL—has been published recently (Zhou et al., 1999).

User guides are available to support the use of MLn (Woodhouse, 1996), MLwinN (Goldstein et al., 1998), HLM (Bryk, Raudenbush, & Congdon, 1996), and VARCL (Longford, 1990), while Singer (1998) discusses fitting hierarchical models using SAS Proc Mixed. The software packages devoted specifically to multilevel or hierarchical modelling have been developed by sophisticated analysts eager to extend the practical applications of these models. This is indicated clearly in their writing, for example, HLM: Bryk and Raudenbush (1992); MLn: Goldstein (1995); and VARCL: Longford (1993). It is safe to say that the perceived user-friendliness of these software programs will depend considerably on the expertise of the user.

We noted in an earlier section that survival analysis can be accomplished using standard statistical software for logistic regressions. The statistical software packages discussed provide the means for carrying out multilevel logistic regression and, therefore, a multilevel survival analysis.

Commentary

The selection of statistical models and methods in this report flows from a belief that the multilevel modelling framework based on random effects is a good one for studying contextual influences on child development, and that developmental studies suited to these methods will become more common in the future. In building these models, we attempted to make them as accessible as possible to the nontechnical reader, stressing potential applications and overlooking issues that might prove distracting or impenetrable. It follows from this that a well-informed statistician should be an integral part of any research team contemplating the use of these methods in their work. Indeed, user understanding of the statistical and mathematical foundations of these methods will have a direct bearing on their successful application. Some of the issues set aside in the main text are identified and discussed briefly in the following commentary.

Modelling Growth with Discrete Response Variables

In this report, the multilevel models described for studying growth as a continuous process are based on continuous response variables with normal distributions.

Of course, the study of growth and change is not always based on continuous response measures: subject responses may be conceptualized and measured as discrete variables. Two types of discrete variables tend to be used in developmental research. One, binary responses are simple classifications usually coded as 0–1 to represent absence or presence of a characteristic. Illness, test success, and school suspension are examples of such variables. Two, experiences or events might be added together to form counts. Examples of these include episodes of bullying in the playground or the number of visits to a hospital emergency room over a specified period of time. Unlike continuous response variables, discrete response variables have non-normal distributions and a restricted range of scores (e.g., are unable to take on negative values).

In developmental studies of growth and change, the statistical modelling of discrete response variables in the form of binary classifications or counts is complicated for several reasons. One, models for discrete response variables in developmental studies are a nonlinear function of the regression coefficients estimated for covariates. This means that modelling discrete response variables using linear regression can be expected to give nonsensible predictions of response. Two, the means and variances of discrete response variables are not separable from one another: they are functionally dependent, unlike the means and variances of continuous response variables.

The general linear model (GLM), which provides the framework for modelling continuous response variables, has been extended in the past 25 years to cover the special challenges posed by discrete response variables. In the context of multilevel modelling, this framework is referred to as the generalized linear mixed model (GLMM) (Hilbe, 1994). It is a "mixed" model in the sense that the model involves a mix of both fixed effects and at least one extra random effect. It is a "generalized linear" model in the sense that the model provides a basis for expressing the associations between covariates and response in linear form. To model discrete responses, the analyst needs to identify an error distribution faithful to the underlying distribution of the response variable and then to choose a *link function* that will allow the covariates to be modelled as linear functions of response. For binary variables, the appropriate transformation is the log odds of response with a binomial error distribution and this is called the logistic regression model. For count variables, the appropriate transformation is the log of response and this is called the Poisson regression model.

The software packages identified earlier provide the means to carry out multilevel modelling with discrete response variables. These packages have been reviewed recently by Zhou et al. (1999), who provide valuable commentary on important features such as data input and management, statistical model capabilities, output, and user friendliness. Although the general approach to modelling discrete response variables extends naturally from continuous responses, the reader should note that some unresolved issues persist in this area.

Sample Size Requirements

A number of very good textbooks have examined issues of statistical power for "traditional" methods (e.g., Cohen, 1988; Kraemer & Thiemann, 1987). In multilevel modelling, sample size estimation for achieving adequate statistical reliability or power is far more complicated and often indeterminable with precision prior to data collection. Only recently have publications appeared on this topic for multilevel modelling (e.g., Hayes & Bennett, 1999; Hedeker, Gibbons, & Waternaux, 1999; Maxwell, 1998; Rochon, 1998; Snijders & Bosker, 1993) and survival analysis involving multiple groups and nonproportional hazards (Ahnn & Anderson, 1998). These papers serve as useful references because they draw attention to the types of information required for determining sample sizes in a primary study. To our knowledge, however, there are currently no general-purpose power tables or software programs that are specific to multilevel models, largely because of additional factors that need to be accounted for in these estimates. For example, statistical reliability or power in multilevel modelling is a function of the standard errors associated with the regression coefficients or βs. Some additional factors that bear on the magnitude of these standard errors include: (1) the extent to which model assumptions, particularly the distribution of the error or residual terms, are met; (2) the extent to which responses are correlated within levels; (3) the mathematical procedure used to estimate variability of response; and (4) the number of observations, individuals, and groupings in the model (Kreft & De Leeuw, 1998, pp. 119–124). Much more work needs to be done to improve our understanding and knowledge about issues of statistical power in multilevel modelling.

Other Modelling Approaches for Longitudinal Data

In the study of child development as a continuous process, the multilevel modelling approach presented in this report draws from a linear random effects model. This model assumes that response is a linear function of the covariates selected for study and that the regression coefficients (i.e., growth trajectories) used to predict response vary from one individual to the next (i.e., exhibit natural heterogeneity) because of random processes. In the section immediately preceding this one, the linear random effects model was extended to the nonlinear case for discrete response variables.

Random effects models are well suited to most studies of child development. In these studies, researchers are interested in modelling between-child variations in rates of growth or change. It is not unreasonable to assume that the heterogeneity in response observed in a sample of children is derived from a distribution that characterises a larger population. The notion that the growth curves of individual children in a development study have been sampled (randomly) from a larger population extends upwards to naturally formed groupings. In a developmental study of children sampled from neighbourhoods, the regression coefficients modelled at this level constitute a sample drawn (randomly) from a larger population of neighbourhoods.

Other models have been developed to estimate regression parameters correctly in longitudinal studies where response variables are correlated. These models go by the name of marginal effects models and conditional or transitional effects models. The analytical objectives and estimating procedures developed for these two types of models differ from each other and differ, in turn, from random effects models. In practical terms, however, the interpretation of the regression coefficients describing population averages are similar for these three models when continuous response variables are the focus of analysis. However, this is not the case when discrete response variables are the focus of analysis. The discussion of these two models below is necessarily brief. Readers interested in learning more about random effects models, marginal models, and conditional or transitional effects models (and possessing some technical expertise) should examine the work of Diggle, Liang, and Zeger (1994), who present a comprehensive overview of statistical models and methods for analyzing longitudinal data.

Marginal effects models. Marginal effects models are so called because the marginal distributions of the response variables are regressed on the covariates chosen for study. The marginal distribution of response refers to the average response for all individuals who share the same value for the covariate. Basically, this is the same type of estimate that is derived in a cross-sectional study. For a binary response variable and a dichotomous covariate, the regression coefficient is the prevalence odds-ratio often used in epidemiological surveys to quantify the strength of association between disorder or disease and some independent characteristic of individuals. The predicted values generated in these models are assumed not to depend on the error structure characterizing within-individual responses. Accordingly, the within-individual correlation structure associated with response is modelled independently from the regression equation that predicts response based on the covariates.

Liang and Zeger (1986) developed generalized estimating equations (GEE) to account for correlations between discrete variables in generalized linear regression models used to estimate marginal effects. GEEs were developed primarily for epidemiological applications where investigators are interested in estimating the prevalence or incidence of disease. Software for implementing GEEs is available in a variety of all-purpose statistical packages including SAS, Stata, Sudaan, and S-Plus. Horton and Lipsitz (1999) have provided a review of these particular software programs.

Conditional effects models. Conditional effects models (also referred to as transitional or Markov models) are so called because the conditional mean (i.e., the average response that is conditional on the immediately preceding response) is regressed on the covariates selected for study. Thus, the individual's current response is influenced by his or her pattern of previous responses. Conditional effects models that are conditional only on the previous response are often called first-order autoregressive models or AR(1) models. Conditional effects models that are conditional on the previous two responses are called second-order autoregressive models or AR(2) models, and so on. Accordingly, the predicted values in conditional effects models do depend on the distribution of responses for the individual, and the error structure arising from within-individual responses must be specified exactly.

Econometricians use conditional models for prediction purposes. This is accomplished by including the previous response with the set of covariates used to estimate the current response—an AR(1) model. Epidemiologists also use conditional models when they wish to estimate the incidence of disease or disorder. Basically, the regression coefficient estimated in a conditional effects model is the same as one derived in an incidence study: for a binary response and a dichotomous covariate, this would be the incidence odds ratio. Transitional models can also be fitted using GEE (Diggle et al., 1994, p. 145).

More Recent Developments in Modelling Development

Covariance structure analysis. In recent years, attention has focused on adapting covariance structure analysis for modelling individual growth and development (Muthen & Curran, 1997; Willett & Sayer, 1994). Covariance structure analysis is a two-step procedure which involves the construction of latent variables (hypothesized constructs or factors derived from mathematical equations applied to observed responses and a specification of their interrelationships (a structural model that estimates the direction and strength of association between the latent variables). It is a very powerful analytic tool that offers considerable flexibility in model specification and explicitly distinguishes between true variation of measurement (the latent variable) and measurement error (unique variation associated with the observed responses making up the latent variable). Although relatively rare, applications of these methods are starting to appear in the published literature (e.g., Kowalski-Jones & Duncan, 1999; Sayer & Willett, 1998). As experience is gained in the use of these methods, their similarities and differences with the models presented in this article will become clearer.

Multivariate multilevel modelling. There are also some variants of multilevel modelling that will lead to exciting applications for those studying developmental processes. One is a multivariate multilevel model, whereby two or more outcome variables are modelled simultaneously (Goldstein, 1995; Thum, 1997). For example, the onset of puberty appears to be related to a girl's body mass index (Brundtland, Liestol, & Walloe, 1980), but there is considerable speculation about the causal mechanisms. It may be that some psychological problems, such as internalized behaviour disorders, cause both overweight and early puberty. A multivariate multilevel model would allow the researcher to simultaneously model BMI and pubertal development as outcome variables over the adolescent period, with measures of behaviour disorders as covariates.

Another promising approach is the integration of the techniques of geographers into developmental models. In all of the models presented above, we treat individuals as independent entities, without attention to the course of development of other individuals in their immediate vicinity. The same applies to the multilevel models; even though they account for correlations among individuals in the same groups, the groups are treated as independent entities. Geographers have developed models for taking account of the dependence among geographically contiguous units, called "spatial auto-correlation" (see Fotheringham & Charlton, 1994), and for estimating regression analyses "locally" rather than "globally" (Fotheringham, Charlton, & Brunsdon, 1997). At the beginning of this paper, we indicated that great strides in developmental research have been occurring as researchers begin to study the role of factors related to the contexts in which people live and work. Bringing geography to bear on this avenue of research promises to advance the work significantly. It will allow developmentalists to understand the nature of variation in developmental trajectories within and among local communities, thereby paving the way for stronger research designs which focus data collection in areas where there are pronounced differences among local communities.

Concluding Remarks

In this report, we have attempted to give nontechnical readers some insight into how a multilevel modelling framework can be used in longitudinal studies to assess contextual influences on child development when study samples arise from naturally formed groupings. Viewing child development as a continuous process, defined as the acquisition and loss of functional characteristics measured by continuous response variables, focused attention on growth curve analysis as a special case of multilevel modelling. Further, it was shown how this framework could exploit the relational structure of hierarchical data existent in naturally formed groupings to test substantive hypotheses about contextual influences on development. Viewing child development as a transitional process, defined as the occurrence and timing of major life experiences and transition points, attention focused on discrete-time survival analysis. Once again it was shown how a multilevel modelling framework could be used to address research questions about contextual influences on this type of developmental phenomenon.

New statistical approaches provide opportunities to increase our knowledge about the world by making it possible to address more informative questions. It is not at all necessary for busy clinicians to master these approaches in order to advance their research. Indeed, the theoretical and mathematical bases for these methods are difficult to grasp, and their effective use will most likely require a seasoned analyst. It is a *sine qua non*, however, to know about these methods if they are to have a chance to strengthen developmental research. In professing the usefulness of growth curve modelling and discrete-time survival analysis for analyzing developmental processes, we hope that readers new to these methods are able to visualize the possibility of using them to advance their work.

Acknowledgements—Dr Boyle was supported by a Canadian Institutes of Health Research Scientist Award. Dr Willms received support from the Canadian Institute for Advanced Research and the U.S. Spencer Foundation. The authors acknowledge gratefully the helpful criticisms, comments and suggestions provided by the three anonymous reviewers of the manuscript.

References

Angold, A., Erkanli, A, Costello, E. J, & Rutter, M. (1996). Precision, reliability and accuracy in dating of symptom onsets in child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry*, *37*, 657–664.

Ahnn, S., & Anderson, S. J. (1998). Sample size determination in complex clinical trials comparing more than two groups for survival endpoints. *Statistics in Medicine*, *17*, 2525–2534.

Boyce, W. T., Frank, E., Jensen, P. S., Kessler, R. C., Nelson, C. A., Steinberg, L., & the MacArthur Foundation Research Network on Psychopathology and Development. (1998). Social context in developmental psychopathology: Recommendations for future research from the MacArthur Network on psychopathology and development. *Developmental Psychopathology*, *10*, 143–164.

Boyle, M. H. (1995). Sampling in epidemiological studies. In F. C. Verhulst & H. M. Koot (Eds.), *The epidemiology of child and adolescent psychopathology* (pp. 337–365). Oxford: Oxford University Press.

Boyle, M. H., Cunningham, C. E., Heale, J., Hundert, J., McDonald, J., Offord, D. R., & Racine, Y. A. (1999). Helping children adjust—A Tri-Ministry Study: I. Evaluation methodology. *Journal of Child Psychology and Psychiatry*, *40*, 1051–1060.

Brundtland, G. H, Liestol, K., & Walloe, L. (1980). Height, weight and menarcheal age of Oslo schoolchildren during the last 60 years. *Annals of Human Biology*, *7*, 307–322.

Bryk, A. S. (1980). Analyzing data from premeasure/postmeasure designs. In S. Anderson, A. Auquier, W. W. Hauck, D. Oakes, W. Vandaele, & H. I. Weisberg (Eds.), *Statistical methods for comparative studies* (pp. 235–260). New York: Wiley.

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: Applications and data analysis methods.* (*Advanced Quantitative Techniques in the Social Sciences*). Newbury Park, CA: Sage.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *HLM. Hierarchical linear and nonlinear modelling with the HLM/2L and HLM/3L programs.* Chicago: Scientific Software International.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Conduct Problems Prevention Research Group. (1999 a). Initial impact of the Fast Track prevention trial for conduct problems: I. The high risk sample. *Journal of Consulting and Clinical Psychology*, *67*, 631–647.

Conduct Problems Prevention Research Group. (1999 b). Initial impact of the Fast Track prevention trial for conduct problems: II. Classroom effects. *Journal of Consulting and Clinical Psychology*, *67*, 648–657.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society Series B*, *34*, 187–202.

Cox, D. R., & Oakes, D. (1984). *Analysis of survival data.* London: Chapman & Hall.

Diggle, P. J., Liang, K.-Y., & Zeger, S. L. (1994). *Analysis of longitudinal data.* Oxford: Clarendon Press.

Fotheringham, A. S., & Charlton, M. (1994). GIS and exploratory spatial data analysis: An overview of some recent research issues. *Geographical Systems*, *1*, 315–327.

Fotheringham, A. S., Charlton, M., & Brunsdon, C. (1997). Measuring spatial variations in relationships with geographically weighted regression. In M. M. Fischer & A. Getis (Eds.), *Recent developments in spatial analysis.* Heidelberg, Germany: Springer-Verlag.

Francis, D. J., Fletcher, J. M., Stuebing, K. K., Davidson, K. C., & Thompson, N. M. (1991). Analysis of change: Modeling individual growth. *Journal of Consulting and Clinical Psychology*, *59*, 27–37.

Gibbons, R. D., Hedeker, D., Elkin, I., Waternaux, C., Kraemer, H. C., Greenhouse, J. B., Shea, M. T., Imber, S. D., Sotsky, S. M., & Watkins, J. T. (1993). Some conceptual and statistical issues in the analysis of longitudinal psychiatric data. *Archives of General Psychiatry*, *50*, 739–750.

Goldstein, H. (1995). *Multilevel statistical models.* London: Arnold.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., & Healy, M. (1998). *A user's guide to MlwiN.* London: Institute of Education, University of London.

Greenhouse, J. B., Stangl, D., & Bromberg, J. (1989). An introduction to survival analysis: Statistical methods for analysis of clinical trial data. *Journal of Consulting and Clinical Psychology*, *57*, 536–544.

Hartmann, A., Schulgen, G., Olschewski, M., & Herzog, T. (1997). Modeling psychotherapy outcome as event in time: An application of multistate analysis. *Journal of Consulting and Psychology*, *65*, 262–268.

Hayes, R. J., & Bennett, S. (1999). Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology*, *28*, 319–326.

Hedeker, D., Gibbons, R. D., & Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics*, *24*, 70–93.

Henry, B., Moffit, T. E., Avshalom, C., Langley, J., & Silva, P. A. (1994). On the "Remembrance of Things Past": A longitudinal evaluation of the retrospective method. *Psychological Assessment*, *6*, 92–101.

Hilbe, J. M. (1994). Generalized linear models. *American Statistician*, *48*, 255–265.

Horton, N. J., & Lipsitz, S. R. (1999). Review of software to fit generalized estimating equation regression models. *The American Statistician*, *53*, 160–169.

Hundert, J., Boyle, M. H., Cunningham, C. E., Heale, J., McDonald, J., Offord, D. R., & Racine, Y. A. (1999). Helping children adjust—A Tri-Ministry Study: II. Program effects. *Journal of Child Psychology and Psychiatry*, *40*, 1061–1073.

Huttenlocher, J. E., Haight, W., Bryk, A. S., & Seltzer, M. (1988). *Parental speech and early vocabulary development*. Unpublished manuscript, University of Chicago, Department of Education, Chicago.

Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: Wiley.

Keiding, N. (1999). Event history analysis and inference from observational epidemiology. *Statistics in Medicine*, *18*, 2353–2363.

Koot, H. M. (1995). Longitudinal studies of general population and community samples. In F. C. Verhulst & H. M. Koot (Eds.), *The epidemiology of child and adolescent psychopathology* (pp. 337–365). Oxford: Oxford University Press.

Kowalski-Jones, L., & Duncan, G. J. (1999). The structure of achievement and behavior across middle childhood. *Child Development*, *70*, 930–943.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel models*. London: Sage Publications.

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

Longford, N. T. (1990). *VARCL. Software for variance component analysis of data with nested random effects* (*maximum likelihood*). Princeton, NJ: Educational Testing Service.

Longford, N. T. (1993). *Random coefficient models*. Oxford: Oxford University Press.

Ma, X., & Willms, J. D. (1999). Dropping out of advanced mathematics: How much do students and schools contribute to the problem? *Educational Evaluation and Policy Analysis*, *21*, 365–383.

Manor, O., & Kark, J. D. (1996). A comparative study of four methods for analysing repeated measures data. *Statistics in Medicine*, *15*, 1143–1159.

Maxwell, S. E. (1998). Longitudinal designs in randomized group comparisons: When will intermediate observations increase statistical power? *Psychological Methods*, *3*, 275–290.

Muthen, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, *2*, 371–402.

Miller, R. G. (1981). *Survival analysis*. New York: Wiley.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.

Nich, C., & Carroll, K. (1997). Now you see it, now you don't: A comparison of traditional versus random-effects regression models in the analysis of longitudinal follow-up data from a clinical trial. *Journal of Consulting and Clinical Psychology*, *65*, 252–261.

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analysing repeated measures designs: An extensive primer. *Psychological Bulletin*, *97*, 316–333.

Raudenbush, S. W., & Bryk, A. S. (1988). Methodological advances in analyzing the effects of schools and classrooms on student learning. In E. Z. Rothkopf (Ed.), *Review of research in education*, *Vol. 15* (pp. 423–475). Washington, DC: American Educational Research Association.

Raudenbush, S. W., & Willms, J. D. (Eds.) (1991). *Schools, classrooms, and pupils: International studies of schooling from the multilevel perspective*. New York: Academic Press.

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*, 307–335.

Rochon, J. (1998). Applications of GEE procedures for sample size calculations in repeated measures experiments. *Statistics in Medicine*, *17*, 1643–1658.

Sayer, A. G., & Willett, J. B. (1998). A cross-domain model for growth in adolescent alcohol expectancies. *Multivariate Behavioral Research*, *33*, 509–543.

Schacter, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist*, *54*, 182–203.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, individual growth models. *Journal of Educational and Behavioural Statistics*, *24*, 323–355.

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 timing of events. *Journal of Educational Statistics*, *18*, 155–195.

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18*, 237–259.

Special Surveys Division. (1996). *Special surveys*, *National Longitudinal Survey of Children and Youth: User's handbook and microdata guide*. (*Microdata Documentation* 89*M*0015*EPE*) Ottawa, ON: Statistics Canada and Human Resources Development Canada.

Speer, D. C., & Greenbaum, P. E. (1995). Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. *Journal of Consulting and Clinical Psychology*, *63*, 1044–1048.

Thum, Y. M. (1997). Hierarchical linear models for multivariate outcomes. *Journal of Education and Behavioural Statistics*, *22*, 77–108.

Willett, J. B. (1989). Some results on the reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement*, *49*, 587–602.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin*, *116*, 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 déjà-vu all over again: Using multiple-spell discrete-time survival analysis. *Journal of Educational Statistics*, *20*, 41–67.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology*, *10*, 395–426.

Willms, J. D. (1998). *Assessment strategies for Title I of the Improving America's School Act*. Report prepared for the Committee on Title I Testing and Assessment of the National Academy of Sciences.

Willms, J. D., & Jacobsen, S. (1990). Growth in mathematics skills during the intermediate years: Sex differences and school effects. *International Journal of Educational Research*, *14*, 157–174.

Woodhouse, G. (1996). *Multilevel modelling applications: A guide for users of MLn*. London: Institute of Education, University of London.

Zhou, X.-H., Perkins, A. J., & Hui, S. I. (1999). Comparisons of software packages for general linear multilevel models. *The American Statistician*, *53*, 282–290.

Manuscript accepted 23 August 2000