Volume 13 Number 4 1989 ISSN 0883-0355

International Journal of Educational Research

Research on Teachers' Professional Lives

Michael Huberman
(Guest Editor)

PERGAMON PRESS

Oxford • New York • Beijing • Frankfurt • São Paulo • Sydney • Tokyo • Toronto

Sponsored by the Foundation for Educational Research in The Netherlands (SVO)

CHAPTER 6

TWO TYPES OF QUESTION ABOUT TIME: METHODOLOGICAL ISSUES IN THE ANALYSIS OF TEACHER CAREER PATH DATA

JOHN B. WILLETT and JUDITH D. SINGER

Harvard University, Graduate School of Education, Appian Way, Cambridge, MA 02138, U.S.A.

Abstract

Quantitative researchers studying teacher career paths ask two types of question about time: questions about *duration* (in which time is the outcome) and questions about *change* (in which time is a predictor). Although asymmetrical in their conceptual placement of time, both questions ask what happens to individual teachers and how variation across teachers is related to background, training and environment. In this chapter, we develop a conceptual framework that integrates methods for addressing both types of questions. We argue that traditional *two*-wave methods are fundamentally flawed and that to answer questions about time, researchers must gather *multi*-wave data. We outline new methods that capitalize on the richness of the longitudinal perspective — proportional hazards modeling and growth modeling.

Introduction

Researchers studying teacher career paths typically investigate two types of questions. The first type of question is about time itself. For instance, researchers ask: How long do teachers remain in the profession and how does employment duration vary by teacher characteristics? Do most teachers leave within a few years on the job or do equal proportions leave each year until retirement? Do secondary school teachers stay longer than primary school teachers? Do we lose the best and the brightest?

The second type of question focuses on changes in teachers' attitudes, perceptions and behavior over time. Does classroom performance change with experience? Is the rate of change constant during a teacher's career, or do newer teachers change more rapidly? Does performance stabilize after a while or do teachers continue to change? Do school policies affect these changes?

Although both types of question focus on the variable *time* they differ fundamentally in how time is conceptualized. The first treats *time as an outcome,* a dependent variable —

employment duration is modeled as a function of teacher characteristics and school policies. The second treats *time as a predictor*, an independent variable. Teacher attitudes, behavior and performance are modeled as a function of time and other teacher characteristics.

Both types of question pose serious methodological challenges for applied researchers. For questions of the first type, the difficulty stems from the nature of the outcome. For practical reasons, employment duration data are collected for only a limited number of years; as a result, not all teachers are observed throughout their entire career. For these latter teachers, we do not know the value of the dependent variable (employment duration). Not surprisingly, researchers have had difficulty developing methods for analyzing an outcome that is by its very nature *missing* for many people under study!

For questions of the second type, the problem is not missing data, but what to do with all the data that *are* available. How can data collected at two or more points in time be used to tell us about change? For decades, methodologists (incorrectly) convinced themselves that good measures of change were a 'holy grail', often sought, but never found. Difference scores, intuitively appealing measures of change, were denounced as unreliable and invalid. Many alternatives proposed as superior (such as the residualized gain score) were actually inferior. The heated and seemingly unresolvable debate led some methodologists to suggest that researchers should not attempt to study change at all!

Fortunately, recent methodological advances have led to sound strategies for addressing both types of question. For questions of the first type, techniques adapted from biostatistics — the methods of survival analysis — can be used to model data in which time is the outcome, even when the outcome, employment duration, is missing for many teachers. For questions of the second type, methodologists have developed better approaches for measuring change that capture the intricacies of individual growth and development.

In this chapter, we introduce the traditional approaches to addressing both types of question, identify their flaws, and outline the new methods that resolve these problems. We have opted for simplicity over completeness; our goal is to *introduce* the methods, not to provide all the documentation necessary for implementing the requisite analyses. We provide references to technical papers for readers interested in pursuing methodological details.

Time as an Outcome

Quantitative studies of *time as an outcome* typically consist of two phases. In the first phase, researchers study *individual employment patterns* — how long do teachers stay on the job? Does the probability of leaving increase, decrease or remain constant over time? If it changes, does it change at a steady rate, or does it level out after a few years? In the second phase, researchers study *between-individual differences in employment duration* — how does length of employment differ across teachers? Do some teachers stay longer than others? Are they the better paid teachers?

Many researchers studying these questions have followed samples of teachers for only two years and simply estimated the proportion of teachers who remained at the end of this time. But as we show below, these 'two-wave' studies are inherently inferior to longer studies (in which teachers are tracked for many years) because the former cannot capture

the full richness of the employment history. To understand why some teachers stay longer than others, and to identify who is most likely to leave, samples of teachers must be followed for *extended* periods of time. Multi-wave studies generate more exploitable information, thereby allowing more powerful statistical analyses. It is this transition, between two-wave and multi-wave studies, that has been the genesis of the superior methodology for analyzing teacher career path data.

Traditional Approaches

Two-wave studies

The traditional way of investigating teacher employment duration has been to compare a pair of cross-sectional surveys of teachers employed in particular districts or geographic areas (see, e.g., Grissmer & Kirby, 1987). To examine first-phase questions of individual teacher employment patterns, teachers in the two surveys are matched and attrition rates (the proportion of teachers present in the first year, who left by the second) are estimated. To examine second-phase questions of between-teacher differences in employment duration, the sample is subdivided into strata — males and females, primary- and secondary-school teachers — and attrition rates are re-estimated within each group separately.

The major contribution of two-wave studies has been to document a pattern familiar to school administrators: Attrition is a U-shaped function of experience — high among new teachers, low for many years among experienced teachers, and high as teachers near retirement. But, when attempting to identify the correlates of duration, researchers with only two waves of data have been severely restricted. Within the two-year period studied, most teachers will not have left their jobs. So although researchers may have complete data for all teachers on many variables of interest, the value of the outcome (employment duration) is unknown for many teachers. For these teachers, all a researcher knows is that duration exceeds a specific value — the length of data collection. In technical terms, these teachers have *right-censored* employment durations.

What strategies have researchers used to analyze right-censored data? Some have tried to circumvent the problem by studying only those teachers who left before data collection ended. But this strategy is conceptually flawed because it necessarily leads to an *underestimate* of the true average length of service. The very existence of continuing teachers establishes that the true average must be longer than average observed among those who left. *Continuing teachers tell us a lot about the probability that teachers stay longer than the data collection period.*

Recognizing this problem, other researchers searched for methods that would allow them to incorporate into their analyses data for teachers who were still employed at the end of data collection. One popular approach was (and still is) to dichotomize duration at an arbitrary time point, usually the length of data collection. Everyone who left during data collection is assigned a value of 1; everyone who remained for the entire data collection period a value of 0. This new dichotomous outcomes variable was regressed upon predictors that described the background, training and environment of the sampled teachers, in an attempt to identify key predictors of employment duration.

Although dichotomization is a useful exploratory strategy, its routine application is

unsatisfactory for two related reasons. First, the choice of a time point at which to dichotomize is just too arbitrary. Most researchers choose the length of data collection, but this time point usually has little substantive meaning. This is especially problematic because empirical results can be very sensitive to the cut point. Second, dichotomization sets aside potentially important information — all the variation in employment duration on either side of the cut point. If the cut point is 8 years, for example, teachers who leave after 7 years are equated with teachers who resign after only 1. This decrease in variation leads to a decrease in statistical power.

Multi-wave studies

The longer the length of data collection, the less serious the censoring problem. As a result, multi-wave studies have fewer censored observations and more information about employment history. Indeed, as early as 1965, Whitener realized that two-waves of data could not capture the complexity of teacher career paths. To understand teacher career patterns more fully, she followed 937 teachers who began teaching between 1951 and 1953 and gathered ten years of data, finding that: (a) the probability of leaving a district was high immediately after entry, but diminished over time, leveling off approximately four to five years later; and (b) the length of employment differed as a function of the age and gender of the entering teachers. Similary, Charters (1970) followed 2,064 teachers for 4 years, Schlechty and Vance (1981) followed 32,131 teachers for 1 to 7 years, and Mark and Anderson (1985) followed 14,827 teachers for 1 to 13 years. Although more sophisticated analytic methods are now available for analyzing such longitudinal career path data, these researchers should be complimented for this crucial reconceptualization.

To examine career persistence, these researchers constructed *life tables,* which include (among other statistics) information on the proportion of teachers in the sample 'surviving' from one year to the next — the survival probability. Figure 6.1 displays a series of sample survival probabilities for over 14,000 teachers who began teaching in the St. Louis Metropolitan area between 1969 and 1982 (Singer & Willett, 1988).

A collection of survival probabilities like Figure 6.1 is known as an estimated survivor *function.* The population survivor function, denoted by $S(t)$, specifies the population probability that a randomly selected teacher will remain employed beyond time t, for all t:

$$S(t) = \text{Prob}[\text{survival beyond } t]. \tag{6.1}$$

So, the survivor function is simply a list of probabilities — the probability that a teacher will teach for more than 1 year, 2 years, 3 years, and so on. At the beginning of a study, when all teachers have just begun their jobs, 100% of them are teaching, and $S(0) = 1.00$. After 1 year, only 77.7% are still teaching, a decrease of 22.3 percentage points; after 2 years, 66.4% are still teaching, a further decrease of 11.3 percentage points. As time passes, teachers gradually leave, and the survival function monotonically decreases, eventually leveling out. Because some teachers do not leave before the end of data collection, there are right-censored observations, and the survivor function never reaches zero.

To investigate second phase questions of between-teacher differences in employment duration, estimated survival functions and survival plots can be constructed separately for subgroups of teachers. For instance, Mark and Anderson (1978, 1985) estimated survival functions separately by teacher's year of entry and gender; Charters (1970) examined

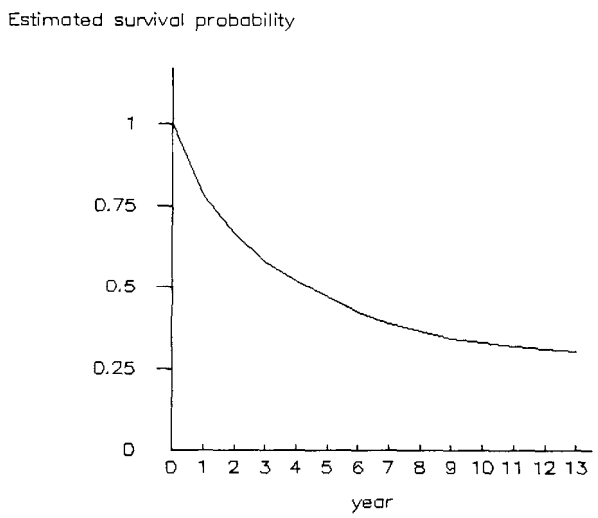

Figure 6.1 Estimated survival function for 14,000 teachers who began teaching in the St. Louis Metropolitan area between 1969 and 1982.

survival as a function of age, gender, teaching level, prior teaching experience, district size and district wealth.

Although reformulating the question of employment duration in terms of the survivor function was a great leap forward beyond simple two-wave attrition analyses, many researchers in this area did not posit an appropriate statistical model relating employment duration to teacher characteristics. Instead they relied solely upon an *ad hoc* jumble of exploratory and nonparametric tools for comparing the survivor functions of different groups. Although the exploratory analyses should always precede formal statistical inference, the adoption of a coherent analytic framework will always lead to superior analyses. Recent methodological improvements provide just such a framework.

Modern Approaches

In recent years, methodologists have developed better methods for analyzing 'time as the outcome'. These new methods incorporate much more of the information present in right-censored data. Known as *survival analysis* (Cox & Oakes, 1984; Kalbfleish & Prentice, 1980; Miller, 1981) or *event-history analysis* (Allison, 1984; Tuma, 1982) these methods do not model duration or the survivor function directly, but a mathematical reexpression that remains meaningful in the presence of right-censoring — the *hazard function*.

To understand why the hazard function, and not the survivor function, is the basis of survival analysis, an inherent problem with the survivor function must be recognized. Teachers can only leave teaching in a given year if they have already survived the preceding year(s). As a result, the survivor function for year t confounds cumulative information on survival for all preceding $t-1$ year with specific information on survival in year t. Thus, the hazard function was defined in order to describe the 'risk' of leaving in *any*

particular year, given survival up to that time. Just as the survivor function is a list of probabilities, one for each year, the hazard function is a list of 'risks'.

When employment duration is measured in years (so that its values are discrete, not continuous), these 'risks' are actually conditional probabilities. Specifically, the hazard function evaluated in year t is the *population probability that a teacher will leave teaching that year, given that she has survived until the beginning of the year:*

$$h(t) = \text{Prob}[\text{leaving between } t \text{ and } t+1 | \text{survival until } t]. \quad (6.2)$$

The elements of the hazard function are the probabilities of leaving teaching in each year *conditional* on having remained through the end of the previous year. It indicates, for instance, whether the second year of teaching is particularly risky or whether the third year is less risky than the second, and so on.

Figure 6.2 displays the hazard function corresponding to the survivor function in Figure 6.1. It is highest in the earliest years, when many new teachers leave, and then it levels off as the survivor function levels off. This correspondence illustrates the link between these two functions. *The hazard function measures how rapidly the slope of the survivor function decreases.* If the survivor curve drops dramatically, many teachers have left their jobs suddenly and hazard is elevated in the corresponding year.

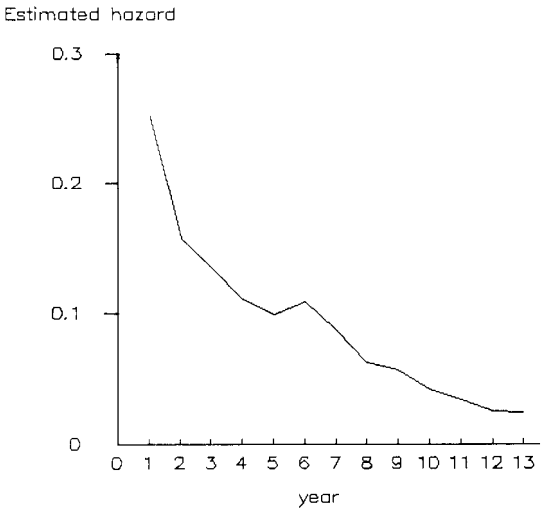

Figure 6.2 Estimated hazard function for 14,000 teachers who began teaching in the St. Louis Metropolitan area between 1969 and 1982.

Having defined a new function which appropriately captures teachers' career mortality, how should second-phase questions about between-teacher differences in career persistence be addressed? For example, how can we find out if career paths differ by teacher subject area, training, or other characteristics? The obvious solution is to explore the relationship between this new summary of teacher career patterns — the hazard function — and the selected teacher characteristics. To do this, hazard is simply expressed as a function of gender, training and so forth in much the same way as a regression model represents the relationship between a (noncensored) outcome and any number of well-chosen predictors.

But what type of 'hazard model' is the most appropriate? Although a variety of models have been proposed (see, e.g., Allison, 1984), one of the simplest, most popular and most robust was suggested by Cox (1972). In the Cox model, hazard is expressed as an exponential function of the predictors. With a single predictor X_p, where p represents the p_{th} person, we write:

$$\log_e[h_p(t)] = \log_e[h_0(t)] + \beta X_p \tag{6.3}$$

where $h_0(t)$ is an unknown 'baseline' hazard (the hazard function for teachers for whom $X_p = 0$) and β is an unknown population parameter (similar to an ordinary regression coefficient) to be estimated. If β is positive, larger values of X are associated with higher hazard; if β is negative, larger values of X are associated with lower hazard; if β is near zero, then X is unrelated to hazard. Just as additional predictors can added to a simple linear regression model, so can further terms be added to the right hand side of (6.3), in which case log-hazard is expressed as a linear combination of the predictors.

Equation (6.3) is known as a *proportional hazards model*. The most popular strategy for estimating its coefficients — the β's — is the method of partial-likelihood, often called Cox regression (Cox, 1972). Although proportional hazards models are based upon complex statistical theory, and most of the relevant papers assume a high degree of mathematical sophistication, these models are now being used more frequently by empirical researchers largely because they have become available in popular statistical packages such as SAS and BMDP (see, for example, Murnane, Singer, & Willett, 1988).

Just like its cousin, maximum likelihood estimation, the method of partial-likelihood yields point estimates of the β's, standard errors and associated hypothesis tests. The improvement in prediction associated with a single predictor can be examined, as can the effect of adding several predictors simultaneously, using a strategy similar to increment-to-R^2 testing in multiple regression. For the fitting process to be valid, several key assumptions must be met. Willett and Singer (1988) discuss exploratory methods that can be used to assess the tenability of these assumptions and present strategies for doing high quality data analysis with these models.

The proportional hazards model in (6.3) expresses hazard for the p_{th} teacher as a sum of two terms: one depending *solely upon time* that is identical for all teachers — $\log_e[h_0(t)]$ — and the other *unrelated to time* that is a function of the predictor — βX_p. Under this parameterization, the logarithm of the baseline hazard function shifts vertically by the amount β for each unit difference in X. This means that the *shape* of the log baseline hazard is identical for all teachers, but it is shifted up or down depending upon the teacher's values of X.

Having estimated the unknown parameters in equation (6.3), the survivor function can be reconstructed from the fitted hazard function. Most statistical packages do this automatically, producing fitted survivor plots for different values of X. For example, Figure 6.3 displays the fitted survival plots for all white female teachers under age 30 who began teaching in Michigan in either 1972 or 1973 (Murnane *et al.*, 1988). The several fitted functions indicate differences in survival among teachers of different academic subjects.

Time as a Predictor

Quantitative studies of *time as a predictor* also consist of two phases. In the first phase,

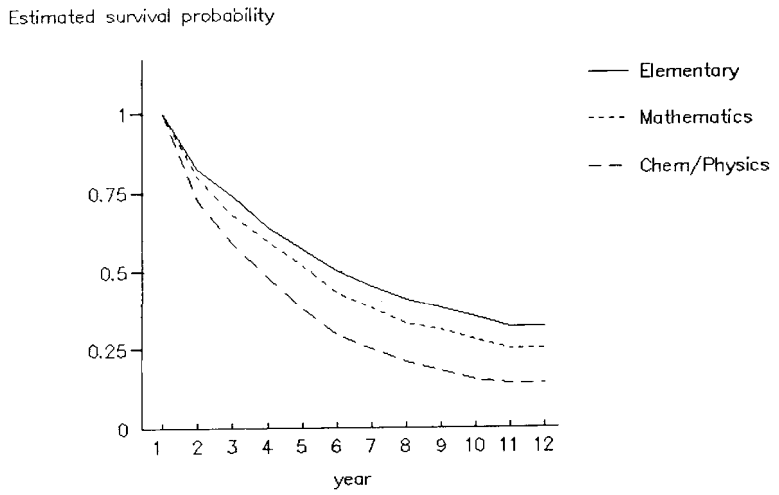

Figure 6.3 Fitted survival functions for all white female teachers under age 30 who began teaching in Michigan in either 1972 or 1973 and who taught in elementary schools or who taught high school mathematics and chemistry/physics.

researchers study *individual growth* — how do teachers change over time? Does job satisfaction, say, increase, decrease or remain constant? Does it grow at a steady rate, or does it level out after a few years? In the second phase, researchers study *between-individual differences in growth* — how do the changes differ across teachers. Do different teachers change at different rates? Do secondary school teachers become satisfied more rapidly than primary school teachers? And so on.

The measurement of individual change is central to both the first and second-phase analyses. But some methodologists have incorrectly argued that change could not be measured accurately. Their mistake was to conceive of change too narrowly as an *increment* — the difference occurring between two time points. They argued that between pre- and post-measurement, each teacher acquired a certain amount of satisfaction and that the goal was to determine the size of this increment. Because it seemed to be difficult, and sometimes impossible, to measure the increment well, many methodologists decided that change could not be measured at all.

In the past decade, methodologists have shown that change should not be conceptualized as an increment, but as a *continuous process of development.* As such, change must be studied over time, and adequate change measurement requires data at more than two time-points.

Traditional Approaches

Many researchers studying change observe the status (say, job satisfaction) of teachers at two points in time. Then, in the first phase of their analyses, these researchers construct measures of change summarizing each teacher's 'growth'. In the second phase, they study the variation in these summaries across teachers to determine how change is related to other teacher characteristics.

Difference scores

How should growth summaries be computed with only two waves of data? The simplest growth summary is the difference between initial status and final status, the so-called *difference, change,* or *gain* score. The difference score is unbiased, easy to compute and intuitively appealing. Although once highly favoured, it was lambasted through the 1950s, 60s and 70s because of its purported unreliability and (usually negative) correlation with initial status (Bereiter, 1963; Linn & Slinde, 1977). But these criticisms were based on flawed assumptions, and the difference score, and some modifications of it, are now seen as the best you can do with only two waves of data (Rogosa, Brandt, & Zimowski, 1982; Rogosa & Willett, 1983; 1985; Willett, 1988).

Central to arguments for and against the difference score is the distinction between *observed* and *true scores*. Measurement error contaminates all empirical measurement and because of it, observed scores are fallible indicators of true scores. When measuring job satisfaction, for example, you do not record *true* satisfaction, but a combination of true satisfaction and whatever error happens to crop up during the instrument's administration. In mathematical terms, the observed status of person p at time i (X_{ip}) is the sum of his or her true status (ξ_{ip}) and a random measurement error (ϵ_{ip}):

$$X_{ip} = \xi_{ip} + \epsilon_{ip}. \tag{6.4}$$

Under classic psychometric theory, the measurement errors ϵ_{ip} are drawn independently from identical normal distributions with zero mean and constant variance.

The problem is that interest centers not on the fallible *observed scores,* but on the underlying *true scores*. Observed scores are simply a lens through which we hope to discern the hidden nature of true scores, and ultimately true changes. When measurement errors are small relative to true scores, observed scores closely mirror true scores; the observed ranking of people on X_{ip} closely parallels the true ranking on ξ_{ip}. When measurement errors are large relative to true scores, the data are little more than noise. The signal is disturbed, less intelligible and informative; observed scores may not mirror true scores. so not only does every observed ranking differ from every other observed one, the one you got this time might bear little resemblance to the true one!

To quantify the relative influence of measurement error and true score, methodologists determine an instrument's reliability. Conceptually, reliability captures the *consistency* of duplicate measures of the same characteristic on a group of people; mathematically, it is defined as the ratio of variance of true scores to the variance of observed scores over people. If measurement error variation is small, the true score variance is close to the observed score variance, and reliability is high. If measurement error variation is large, the true score variance is much smaller than the observed score variance (which has been inflated by measurement error), and reliability is low.

Although high reliability is clearly desirable, many of the arguments concerning the difference score fail to distinguish the *two* ways in which observed scores can *appear* to be unreliable. The most important factor is, of course, measurement error — the greater the random error, the lower the reliability. But another factor is *the amount of variation in true scores over people*. If true score variation is limited, as happens when most people in the population have similar true scores, reliability will be low even when using an instrument relatively free from measurement error. A low reliability coefficient does not identify the *source* of any unreliability — measurement error or too little true score variation.

The relationship between reliability and variance in true scores had serious consequences when examining difference scores. The *observed* difference score for the p_{th} teacher, D_p, is the difference between the two observed scores ($X_{2p}-X_{1p}$). Similarly, the *true* difference score is $\Delta_p = (\xi_{2p}-\xi_{1p})$. Using equation (6.4), we see that the observed difference score D_p is the sum of underlying true change, Δ_p, and the difference between the two measurement errors ($\epsilon_{2p}-\epsilon_{1p}$). Δ_p, not D_p, is the real focus of interest; D_p is inflated or deflated by the presence of measurement error.

Some authors have argued that difference scores are always unreliable; others have argued that they are sometimes reliable, but that when they are, they are not valid. These misconceptions arise from misinterpreting the pre-test–post-test correlation as an index of construct validity in expressions for the reliability of the difference score (Rogosa *et al.*, 1982). In many situations, the difference score's reliability is respectable (Rogosa & Willett, 1983) and when differences in true growth are large, the reliability of the difference score can be greater than the reliabilities of the constituent measures (Willett, 1988).

Even if the difference score was unreliable, we still would not necessarily have a problem. Low reliability is not always due to imprecise measurement; it can be due to limited variability. A difference score can be unreliable simply because everyone in the population is growing at approximately the same rate. So although you may know very accurately that everyone changed by ten points (say), people are not differentiated by their growth and so the difference score *appears* unreliable. Low reliability does not always indict the difference-score; sometimes it indicts the policy of using reliability as the sole indicator of measurement quality.

The difference score has also been faulted for its correlation with initial status. Critics claim that if a measure of change is related to initial status, the measure is unfair because it gives "an advantage to persons with certain values of the pretest" (Linn & Slinde, 1977, p. 125). But why *should* growth and status be unrelated? Their connection is an inevitable consequence of growth history. People growing rapidly will have a higher status on later occasions than those growing slowly. Current status is a direct consequence of past growth, past status a mediator of future growth. Growth and status *should* be related.

Others have criticized the difference score not because of its correlation with initial status, but because the sign and size of the correlation differs widely across studies — sometimes positive, sometimes zero, and sometimes negative (see, for instance, Bloom, 1964). But *should* we expect a single value for this correlation? If everyone isn't growing in parallel, the correlation *must* differ depending upon the time chosen as the initial time point. Unless a substantive justification for a 'real' initial time point can be given (say, first year of teaching), why should there be a unique answer to the question: What is the correlation between growth and initial status? Researchers *should* find different values for this correlation (Rogosa & Willett, 1985).

Still others have criticized the difference score because its correlation with initial status is usually negative. Even if this were true, (which it is not — see, for instance, Thorndike (1966)), this still would not be a blanket condemnation. Some researchers have actually created the problem themselves by 'standardizing' the pre- and post-measures to a common variance before computing difference scores (an ill-advised process that destroys information — see Willett, 1988). Others confuse the correlation between observed initial status and the observed difference score with the correlation between *true* measures of status and difference. Unfortunately, the observed estimator is negatively biased due to

the presence (with opposite sign) of the t_1 measurement errors in both observed initial status and the difference score, thus yielding a negative estimate even when the underlying true correlation is positive. We should not reject an unbiased estimate of change — the difference score — simply because an estimator of the association between true change and true initial status is biased (and anyway, this latter bias is easily corrected — see Zeive, 1940).

Although the difference score is not the pariah many critics claim, several improvements are possible, such as Webster and Bereiter's (1963) reliability-weighted measure of change and Lord (1956) and McNemar's (1958) regression-based estimated true change. These estimators trade unbiasedness for the reduction in mean-squared error. Under broad assumptions, these modifications are weighted linear combinations of the observed difference score for the particular teacher and the average observed difference score over the entire population of teachers, with weights dependent on the difference score's reliability. The weighting schemes emphasize the most trustworthy measurements: a teacher's difference score when it is reliable and the average difference score when the lack of variation in growth across teachers make individual difference scores unreliable. But although these modifications are better estimates of true change, they are so highly correlated with the corresponding difference scores that both measures usually lead to similar conclusions.

Further improvements in change measurement can also be realized if you use an estimate of the difference score's reliability in the second-phase analyses. When second-phase analyses are based on *raw* difference scores, measurement error attenuates the analyses and sample correlations underestimate true correlations. This problem can be avoided by disattenuating the second-phase analyses (Willett, 1988). To do so, additional information — an estimate of the difference score's reliability — must be incorporated. However, the disattenuation is highly sensitive to even minor fluctuations in the size of the reliability estimate. And since the quality of this estimate is often dubious (especially when based on internal-consistency estimates from other sources), serious problems can arise.

Residual change scores: An 'improvement' that fails

Motivated by a desire to obtain measures of change that are uncorrelated with initial status, several methodologists proposed the *residual change score*, residuals from a regression of true final status on true initial status. Methodologists have spent much energy identifying properties of estimators of residual change, but their work has led to considerable controversy. Disagreements surround exactly what is being estimated, how well it is being estimated and how residual change scores can be interpreted; questions of logic and substance have also been raised (Rogosa *et al.*, 1982; Rogosa & Willett, 1985; Willett, 1988). The net result is that residual gain scores have been discredited as measures of change.

Modern Approaches

Two waves of data simply cannot reveal the complex trajectories of individual change. Job satisfaction will probably change nonlinearly over time, for example, yet with two waves of data, you are restricted to using the simplest mathematical model for growth —

a straight line. Just as you would question the validity of a regression model fit to two data-points, so should you mistrust a two-wave measure of growth.

With data collected at three or more time-points, the ability to study change improves dramatically. The first step is to assemble *growth records,* the temporally-ordered observed data for each teacher. Preliminary analysis of these growth records begins with plots, for each teacher, of *empirical growth-trajectories,* graphs of observed scores versus time, with a sketched trend-line. Figure 6.4 presents a sample empirical growth trajectory. Inspection of such trajectories can help determine whether growth is linear or curvilinear and comparisons for these trajectories across groups of teachers may suggest important predictors of growth.

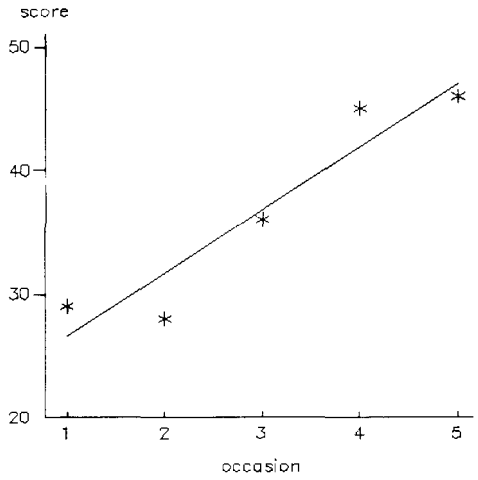

Figure 6.4 Hypothetical straight-line empirical growth trajectory for a single teacher, based on a five-wave individual growth record.

Formal analyses of multi-wave data require the specification of a statistical model to summarize each teacher's trajectory (Bryk, 1977; Rogosa & Willett, 1985). The goal is to select a mathematical function that best represents the observed growth patterns and then fit this model to each teacher's data. Models such as the simple straight-line or the more complex to the nonlinear negative-exponential and quadratic curves are often used. It is even possible to join several different functions together, creating a piecewise growth model. Willett (1988) presents three possible models that have been used to represent growth: the straight-line, the quadratic and the negative-exponential.

How do you select an appropriate individual growth model? Substantive background can sometimes provide a theory to guide model selection; for example, organizational theory might suggest a useful functional form, with increases in satisfaction during early years tapering off over time. Some models constrain scores to rise to a ceiling; others constrain scores to begin from a floor. Such models allow you to test hypotheses not only about the *rate* of growth but also about the *asymptotes.* You might find that one background characteristic of the teacher is related to the ultimate limit on growth and another to the rate of growth.

Once a growth model has been adopted, all teachers are assumed to share this same

shape; teachers are distinguished by values of the constants in the common model. For instance, under the straight-line growth function, the true job satisfaction of all teachers is assumed to be changing linearly with time but different teachers have different initial levels (intercepts) and different rates of growth (slopes):

$$\xi_p(t) = \xi_p(t_*) + \theta_p(t - t_*) \tag{6.4}$$

where the true status of the p_{th} teacher at time t is now represented as $\xi_p(t)$ in order to indicate that true score is a continuous function of time. The intercept $\xi_p(t_*)$ is the true satisfaction of the p_{th} teacher at an arbitrarily selected time t_* (which might be the time at which they began teaching) and the slope θ_p is the rate at which true job satisfaction changes. If the slope θ_p is positive, true satisfaction increases with time; if negative, it decreases.

The intercept and slope for each teacher are called the *individual growth parameters.* Straight lines have two, quadratics and negative-exponentials have three. Although growth models can be selected so that the growth parameters reflect theoretical processes, for parsimony, linear and quadratic functions are most popular. When only a small slice of each person's career is under study or when only a few timepoints are available, a straight line is often satisfactory.

Modeling differences in growth

Teachers with different growth patterns have different growth parameters. If growth is represented by a straight-line, for example, differences between teachers are reflected in differences in their intercepts and slopes.

The individual growth parameters can be used to represent relationships between growth and teacher background characteristics. Suppose that we adopt the simple straight-line model in (6.4) for the growth of each teacher and assume that the rate of growth, θ_p, is systematically related to teacher age at entry, ω_p, over teachers (say). This 'between-teacher' relationship can be represented by the following linear model:

$$\theta_p = \beta_0 + \beta_{\theta\omega}\omega_p + \text{error} \tag{6.5}$$

where θ_p and ω_p are assumed to be linearly related (see Rogosa & Willett, 1985).

The regression parameter $\beta_{\theta\omega}$ summarizes the linear relationship over teachers between true growth rate and teacher age at entry. If older teachers grow more rapidly than younger ones, $\beta_{\theta\omega}$ will be positive; if younger teachers grow more rapidly than older ones, $\beta_{\theta\omega}$ will be negative. Some researchers find it more convenient to speak in terms of the corresponding correlation $\rho_{\theta\omega}$ rather than in terms of the regression coefficient. So a nonzero value of either $\beta_{\theta\omega}$ or $\rho_{\theta\omega}$ indicates that age at entry is linearly related to growth.

A major advantage of multiwave methods is that you are not limited to the simple models presented here. Complex functions can be used to model individual growth and sophisticated models can be used to relate differences in growth parameters to teacher characteristics. For instance, under a negative-exponential growth model, you could adopt a second-phase model describing differences in the *upper limit* — asymptote — of growth or the rate of growth. This perspective opens a new framework for measuring growth, far removed from two waves of data and the difference score.

Fitting the models

If the straight-line growth model has been adopted and you are interested in predictors of the rate of growth, the growth parameters θ_p and the individual regression parameter $\beta_{\theta\omega}$ or correlation coefficient $\rho_{\theta\omega}$ must be estimated. We briefly outline below three strategies for estimating these parameters, each building on the preceding one.

1. *Using ordinary least-squares regression to estimate the individual teacher growth models.* The simplest available approach for handling multi-wave data is to fit a growth model to each teacher's growth record using ordinary least-squares (OLOS) regression — one fitted model per teacher. If the model is appropriate, the obtained slopes and intercepts will be unbiased estimates summarizing the essential information about individual growth (Rao, 1958). These OLS estimates provide measures of growth to be correlated with, or regression upon, teacher characteristics in subsequent second-phase analyses, in much the same way as difference scores were used in two-wave studies. OLS growth-rate estimates are easily obtained and are much more precise measures of change than the difference score.

This strategy offers a powerful and straightforward method for analyzing growth that can be applied on its own or ahead of the more sophisticated strategies described below. But one caveat is in order: Although the OLS estimates are unbiased, they are *fallible* measures of true growth. This fallibility disturbs second-phase analyses and attenuates findings. Willett (1988) presents a method for estimating the magnitude of the measurement error variability from the within-person regression residuals, allowing second-phase analyses to be disattenuated (without requiring information on reliability, as does the two-wave difference score).

2. *Using weighted least-squares regression to estimate the between-teacher regression model.* Second-phase analyses can be improved by incorporating additional information in them. Using an estimate of the precision of each individual growth summary in subsequent analyses not only improves those analyses, but also corrects for the fallibility of the slope estimates as measures of true growth.

To make this correction, the standard errors of the OLS slopes must be estimated for each teacher during the first-phase analyses. These standard errors capture the *precision* with which the individual growth-rates have been estimated and provide a basis for the superior estimation of the second-phase parameters. During second-phase estimation, *weighted* least-squares (WLS) regression is used to fit the model in (6.5), with weights based on the standard errors obtained earlier. Weights of the form $(\text{se}(\theta))^{-2}$ are appropriate for such a procedure because, in the second-phase analysis, the more-precisely determined growth-rates (those with the smallest standard errors) would play the most important role. Willett (1988) reports on work by Hanushek (1974) which suggests another function of the standard errors that provides optimal weights for the second-phase estimation (in the sense that the obtained estimate of $\beta_{\theta\omega}$ is then asymptotically efficient).

The more precise the first phase growth summaries, the more successful the second phase analyses. The best way to increase precision is to collect more waves of data for each teacher. Standard errors of growth rates decrease dramatically as extra waves of data are added. It is for this reason alone that multiwave designs provide more reliable methods for measuring change. *You have complete control over the precision of your findings: the more waves of data, the better.*

3. *Using empirical Bayes estimation to estimate both models simultaneously.*

Methodology for analyzing growth has been advancing very rapidly (Ware, 1985). One sophisticated method of estimating the parameters of the growth models in (6.4) and (6.5) has been provided in a software package called HLM (Bryk, Raudenbush, Seltzer, & Congdon, 1986). Parameter estimates similar to those obtained under strategy 2 above are used as 'start values' for iterative analyses that eventually lead to superior empirical Bayes estimates for both sets of parameters.

Discussion

In this chapter, we have distinguished between two types of question that are typically asked during the investigation of teacher career paths: question about *the duration of teacher employment itself,* and questions about *growth and change in the attributes, behaviors and attitudes of teachers.* Under either class of question, time plays a crucial and central role. But questions of the former type conceptualize time as the outcome ("How long do teachers stay in teaching?") and the latter questions conceptualize time as a predictor ("Do teachers change over time?"). The principal difference between the two classes of question is simply that 'time' has been moved from the left-hand side of the 'equation' to the right-hand side.

Although asymmetrical in their conceptual placement of time, remarkable similarities do exist between the two classes of question. In both cases, research is concerned with what happens to the individual teacher, and how variation from teacher to teacher is related to background, training and environment. Both classes support both *within-teacher* and *between-teacher* subquestions. In the case of 'time as the outcome', we might ask the within-teacher question: How long will this teacher stay in teaching? Or the between-teacher question: Is the employment duration of different types of teacher different? In the case of 'time as a predictor', the corresponding sub-questions are: Is this teacher changing over time? and, Does the growth of different types of teacher differ?

It is the search for *systematic inter-individual differences* in response to the latter subquestions in which there is the most research interest. We want to know whether elementary school teachers remain in the profession longer than secondary school teachers, whether teachers who are paid more stay longer, and so forth. We want to know whether changes in job satisfaction are related to subject-matter speciality, or to gender, or to the socio-economic status of the neighborhood.

Regardless of whether time is treated as an outcome or as a predictor, traditional statistical methods for answering these questions are flawed. It is as difficult to use a 'two-wave' attrition study to investigate differences in career duration between elementary-school and high-school teachers as it is to decide whether job satisfaction is changing more rapidly for men or for women on the basis of a 'two-wave' measure of change. In our chapter, we emphasize that superior methodology has only evolved since researchers began to design longitudinal, multi-wave rather than two-wave, studies.

In the case of studies of employment duration, methodologists also had to address the problem of right-censoring — the problem that, by the end of data-collection, a substantial proportion of the sampled teachers were unlikely to have undergone the event for which the investigator was waiting. They would not yet have left teaching, and therefore the ultimate duration of their employment is unknown. To incorporate information from these censored cases into an eventual analysis of duration, methodologists have

reconceptualized studies of 'time as the predictor' as studies of the hazard function — a function that captures the 'risk' of leaving teaching in any given year. By modeling hazard as a function of selected covariates, questions of systematic inter-individual heterogeneity in employment duration can be answered.

In the case of studies of growth and change, methodologists have moved towards a realization that, if growth is to be measured well, then people must be followed closely over time and measures of individual growth must be based on as many waves of data as possible (and certainly more than two!). With the collection of multi-wave data, changes in teachers over their careers can be carefully modeled and the individual growth parameters used to support second-phase investigations of systematic inter-individual differences in growth. Not only can individual change itself be measured much more successfully but the investigator can enter an arena in which complex hypotheses about the intricacy of individual growth, and how these intricacies are related to characteristics of the teacher's background, can be entertained.

Despite large statistical differences between the methods that have evolved for answering two sorts of question about time, these methods present a common message to the empirical researcher: the more you know, the better off you will ultimately be. If you can collect more information — in this case, longitudinal information gathered carefully over time — and treat it well, then you will necessarily produce fuller, more credible, more accurate answers to your questions.

References

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data,* Sage University Paper Series on Quantitative Applications in the Social Sciences, Series Number 07-046. Beverly Hills, CA: Sage Publications.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in the measurement of change.* Madison, WI: University of Wisconsin Press.

Bloom, B. S. (1964). *Stability and change in human characteristics.* New York: John Wiley.

Bryk, A. S. (1977). An investigation of the effects of alternative statistical adjustment strategies in the analysis of quasiexperimental growth data. Doctoral dissertation. Harvard University Graduate School of Education. *Dissertation Abstracts International,* **38**, 3761B.

Bryk, A. S., Raudenbush, S. W., Seltzer, M., & Congdon, R. J. (1986). *An introduction to HLM: Computer program and users' guide.* University of Chicago.

Charters, W. W., Jr. (1970). Some factors affecting survival in school districts. *American Educational Research Journal,* **7**, 1–27.

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.

Grissmer, D. W., & Kirby, S. N. (1987). *Teacher attrition: The uphill climb to staff the nation's schools.* Santa Monica, CA: The Rand Corporation.

Hanushek, E. A. (1974). Efficient estimators for regressing regression coefficients. *The American Statistician,* **28**, 66–67.

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

Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre- and post-testing periods. *Review of Educational Research,* **47**, 121–150.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement,* **16**, 421–437.

Mark, J. H., & Anderson, B. D. (1978). Teacher survival rates: A current look. *American Educational Research Journal,* **15**, 379–383.

Mark, J. H., & Anderson, B. D. (1985). Teacher survival rates in St. Louis, 1969–1982. *American Educational Research Journal,* **22**, 413–421.

McNemar, Q. (1958). On growth measurement. *Educational and Psychological Measurement,* **18**, 47–55.

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

Murnane, R. J., Singer, J. D., & Willett, J. B. (1988). The career paths of teachers: Implications for teacher supply and methodological lessons for research. *Educational Researcher,* **17**(6), 22–30.
Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics,* **14**, 1–17.
Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin,* **90**, 726–748.
Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement,* **20**, 335–343.
Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika,* **50**, 203–228.
Schlechty, P. C., & Vance, V. S. (1981). Do academically able teachers leave education? The North Carolina case. *Phi Delta Kappan,* **63**, 106–112 (No. 2).
Singer, J. D., & Willett, J. B. (1988). Detecting involuntary layoffs in teacher survival data: The year of leaving dangerously. *Educational Evaluation and Policy Analysis,* **10**(3), 212–224.
Thorndike, R. L. (1966). Intellectual status and intellectual growth. *Journal of Educational Psychology,* **57**, 121–127.
Tuma, N. B. (1982). Nonparametric and partially parametric approaches to event history analysis. In S. Leinhardt (Ed.), *Sociological methodology.* San Francisco, CA: Jossey-Bass.
Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician,* **39**, 95–101.
Webster, H., & Bereiter, C. (1963). The reliability of changes measured by mental test scores. In C. W. Harris (Ed.), *Problems in measuring change.* Madison, WI: University of Wisconsin Press.
Whitener, J. E. (1965). *An actuarial approach to teacher turnover.* Unpublished doctoral dissertation. St. Louis, MO: Washington University.
Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15). Washington, DC: American Educational Research Association.
Willett, J. B., & Singer, J. D. (1988). Doing data analysis with proportional hazards models: Model building, interpretation and diagnosis. Paper presented at the annual meeting of the American Educational Research Association, ERIC Document.
Zeive, L. (1980). Note on the correlation of initial scores with gains. *Journal of Educational Psychology,* **31**, 391–394.

Biographies

John B. Willett and Judith D. Singer are Assistant Professors of Quantitative Methodology, Harvard University Graduate School of Education. Their joint research interests are in the areas of event history and longitudinal analysis, research design, quantitative methodology and the implication of research results for educational policy-making and practice.