Educational Evaluation and Policy Analysis
Winter 1996, Vol. 18, No. 4, pp. 265–283

Methodological Issues in the Design of Longitudinal Research: Principles and Recommendations for a Quantitative Study of Teachers' Careers

Judith D. Singer and John B. Willett
Harvard University

In this paper, we initiate a discussion of the possible methodological features of a potential new national longitudinal study of teachers' careers. We use a review of the substantive and methodological literatures and "pseudolongitudinal" analyses of data drawn from the National Center for Education Statistics' Schools and Staffing Survey and Teacher Follow-Up Survey to support our arguments. Our presentation is framed by six key principles of research design that are then used to support concrete recommendations about whom *should be studied,* how often *they should be observed, and for* how long *the study should continue. Our six design principles assert that, in the new study, we must collect truly longitudinal data, view time as both an outcome and a predictor, collect data on both time-varying and time-invariant measures, collect data prospectively whenever possible, collect data beginning in multiple base years, and collect data at all relevant levels of the organizational hierarchy. Although it is impossible to define a single research design that is optimal for answering all research questions about the teaching career, we recommend that the new study should sample from the population of teachers who are beginning their first, second, third, and subsequent spells, that each of these teachers be followed for at least 12 years (both in and out of teaching), and that they should be measured on at least six occasions during this period. We also recommend that the study be replicated starting in two, if not three, base years. We welcome public comment and discussion of our proposals.*

The National Center for Education Statistics (NCES) is considering conducting a large-scale longitudinal study of teachers' careers. Unlike their current surveys of teachers—the cross-sectional Schools and Staffing Survey (SASS) and the one-year prospective Teacher Follow-Up Survey (TFS)—the new study would follow a sample of teachers over an extended period of time. As would be expected, initial discussions about this study focused primarily on *substantive* issues (see, e.g., Billingsley, 1992; Grissmer & Kirby, 1992a; Theobold & Gritz, 1992; Weiss, 1992). Yet we believe that a discussion of methodological issues is equally important. Should NCES draw a sample from the entire teaching stock or should subsamples of teachers at particular career junctures—say, teachers with 5 and 10 years of experience—be followed? Should data collection continue through retirement or will a shorter period suffice? Should data be collected annually, or is a different interval more appropriate?

The purpose of this paper is to initiate a discussion about the methodological features of a potential new study. We begin by briefly reviewing the study's goals and identifying four major gaps in current knowledge about teachers' careers. This discussion leads us to develop six design principles that we use (in subsequent sections) to support recommendations about the three most crucial design decisions—*whom* should be studied, *how* often should they be contacted, and *how* long should the study continue? We conclude by inviting readers to participate in this debate, using our suggestions as catalysts for discussion with policymakers, government officials, and teacher career researchers.

Research Goals: Why Conduct a New Longitudinal Study of Teachers' Careers?

As quantitative research on teachers' careers has accumulated, its quality has improved. The most generalizable evidence—focused primarily on supply and demand issues—has used data from one of four sources. Administrative records have been used to reconstruct career histories for cohorts of newly hired teachers in individual states (e.g., Grissmer & Kirby, 1992b; Murnane, Singer, & Willett, 1988, 1989; Theobold, 1990). National longitudinal surveys have been used to trace career histories for college graduates who have ever taught in the U.S. (e.g., Hafner & Owings, 1991; Heyns, 1988; Murnane, Singer, Willett, Kemple, & Olsen, 1991). National cross-sectional surveys of teachers have been used to describe previous career decisions and future intentions (e.g., Choy et al., 1993). One-year follow-up surveys tracking national subsamples of teachers have been used to estimate annual attrition rates (e.g., Ingersoll, Han, & Bobbitt, 1995).

Although a review of this literature is beyond the scope of this paper, several findings deserve special mention. Teachers are at greatest risk of leaving in their first few years on the job. Those who score well on standardized tests, have attractive opportunities outside of teaching, or are paid less well are most likely to leave. And teachers are highly mobile: many who leave soon return, making the "reserve" pool of former teachers—not recent college graduates— the major source of supply.

Yet much remains *unknowable* with current data sources. Too few national studies probe why teachers behave the way they do. Too few ask what day-to-day life is really like in schools. Among the many areas in which further work is needed, we believe four warrant priority attention:

- Teacher quality. Despite the central role of teacher quality in debates about education, national studies have yet to assess this elusive construct (see, e.g., Hanushek & Pace, 1995). Although the measurement of teacher quality raises substantive, political, and methodological difficulties that we cannot address here (Kennedy, 1992), this study must tackle these obstacles (Shulman, 1992). After all, attrition may not be a problem if the schools are losing the teachers of lower quality.
- Teachers' work contexts. We have too coarse data about the settings in which teachers teach. The new study must explicitly gather detailed information about administrators, colleagues, and students. Such data would be invaluable for identifying links between teachers' perspectives and those of the administrators they work for, the teachers they work with, and the students they serve; they are also essential for understanding how these linkages change over time.
- Teachers' professional lives. The growing literature on teachers as professionals, leaders, and mentors needs nationally representative data. Do professional development opportunities lead to lower attrition, higher satisfaction, and stronger commitment? How do teachers perceive reform efforts? This new study has the potential to provide generalizable information on work roles, professional development, job commitment, and job satisfaction, allowing researchers to see how these aspects of teachers' lives change over time.
- Teachers' careers. Although quantitative research in this area is fairly sophisticated, much more can be done. No study has juxtaposed career decisions against complete data on preparation, wages, benefits, workplace conditions, and family structure. No study has fully tracked the labor market experiences of former teachers and returning teachers. This new longitudinal study should fill this void.

In highlighting these four areas, we do not mean to imply that they are the *only* ones worthy of investigation. We identify them because they cover substantive issues across several levels of the organizational hierarchy and because each has design ramifications to which we will allude. Unlike a smaller-scale study, which can (and should) be designed to address a targeted set of research questions, a major national undertaking must support analyses of issues across a broad policy spectrum. We refer readers interested in discussing the *substantive* focus of the new study to the commissioned papers referenced in the first paragraph of the paper. Other relevant catalysts for substantive discussion include the teacher career models presented by Billingsley (1993) and the National Academy of Sciences (1992), as well as the excellent qualitative studies of teachers' lives in schools presented in Goodson and Hargreaves (1996), Grossman (1990), Huberman (1989), Johnson (1991), Little and

McLaughlin (1993), and McLaughlin and Oberman (1996).

Implications of the Research Goals: Six General Principles of Longitudinal Design

In contemplating the design features of this study, we identified six principles that we believe underpin the specification of any longitudinal study of teachers' careers. As general principles, they transcend the details, focusing instead on overarching tenets.

Principle Number 1: Collect Truly Longitudinal Data

Ironically, few studies of the teaching career—an inherently longitudinal phenomenon—are truly longitudinal. Most rely upon cross-sectional, retrospective, or two-wave designs. The new study has the potential to usher in a fresh generation of research into teachers' lives. To yield major leaps in knowledge—not just minor additions to what we already know—individual teachers must be contacted repeatedly over an extended period (see, e.g., Duncan & Kalton, 1987).

What's wrong with cross-sectional designs? Simply put, cross-sectional data tell us nothing about change.[1] If cross-sectional data reveal that more experienced teachers have lower levels of professional commitment, we cannot infer that commitment decreases with accumulated experience. Teachers with more experience differ from their less-experienced colleagues in important ways—they entered teaching in different years and they have taught under different working conditions. Observed differences in commitment, then, may be due to nothing more than differences in background characteristics and work experiences, not the number of years the teachers have been on the job.

Two-wave studies are only marginally better. Although they are useful for estimating attrition rates, they, too, are inadequate for studying change. Two-wave designs tell us nothing about the *shape* of each teacher's growth trajectory; they cannot describe how the teacher got from Time 1 to Time 2. Did all the change occur immediately after Time 1, or was progress steady over the entire interval? With two waves of data, we cannot know. The more complex the trajectory, the more waves we need. *Three is the minimum number of waves in any truly longitudinal study.*

Not only must a sufficient number of waves of data be collected, but the study must endure *long* enough to register a sufficient number of changes and transitions. For many of the variables in this study—ranging from pedagogic quality to career transitions—change will be subtle and slow. How rapidly do teachers' teaching skills improve? How soon do teachers leave teaching? With respect to teachers' careers, at least, attrition is near an all-time low (Ingersoll et al., 1995). The study of change in this stable environment must be based on longer longitudinal records than might be tolerable in a period of greater upheaval and foment. As we will soon show, we suggest that this study continue for a minimum of 12 years.

Principle Number 2: View Time As Both an Outcome and a Predictor

Researchers analyzing the data from this study will ultimately conceive of time as either a predictor or as an outcome, the decision depending on the question posed (Willett & Singer, 1989). Those who examine transitions in and out of teaching typically come from an economic tradition that treats time as the conceptual outcome—they examine whether teachers experience transitions and when transitions occur (e.g., Murnane et al., 1991). Those who examine changes in attitudes, knowledge, or behavior over time typically come from sociological and psychological traditions that treat time as a predictor—they study how teacher attributes change over time (e.g., Huberman, 1989). The two types of questions demand different statistical approaches. The former requires methods for analyzing event occurrence, such as *survival analysis* (Singer & Willett, 1991, 1993; Willett & Singer, 1991, 1993); the latter requires methods for analyzing change, such as *individual growth modeling* (Bryk & Raudenbush, 1987; Rogosa, Brandt, & Zimowski, 1982; Willett, 1988).

Study planners usually privilege one of these perspectives, ensuring that the data collection schedule will produce a precise and unbiased summary of either event occurrence or change. Yet the schedule that produces an optimal summary of change may not produce an optimal summary of event occurrence. This multi-purpose study must give each perspective an equal voice, using more frequent fol-

low-ups, for a longer period of time, if necessary. It would be wrong to emphasize one perspective in planning; interested parties from both research traditions deserve data of equal quality.

Principle Number 3: Collect Data on Both Time-Varying and Time-Invariant Measures

All variables can be classified as either *time-invariant* or *time-varying*. Time-invariant variables have values that remain constant over time—ethnicity, college major, year of licensure. Because they never change, they need be measured only once, typically in the base year of data collection. Time-varying variables, in contrast, have values that differ over time—teacher efficacy, work conditions, family composition, to name a few. Data on these variables must be collected repeatedly over time.

Not all time-varying variables fluctuate at the same rate. Teacher efficacy, for example, may change daily while class characteristics may change annually. School characteristics change whenever a teacher moves schools (and sometimes more often). Data collection waves must be spaced closely for variables that fluctuate rapidly. Information on other variables can be collected less regularly.

Whether a time-varying variable will serve as an outcome or a predictor, the observed scores must be *equatable* across occasions of measurement (Goldstein, 1979). Seemingly minor differences across occasions—even those invoked to *improve* data quality—undermine equatability. Between the first SASS and follow-up TFS, for example, the stems and response categories for the items assessing satisfaction with the level of material resources in the school changed rendering responses non-equatable. (Compare the "equivalent" items 29(h) of the 1987–1988 SASS and 27.12 of the 1988–1989 TFS.) At a minimum, item stems and response categories must be the same. Although administering an identical instrument repeatedly is not without problem, these issues pale when compared with the consequences of measurement modification (Light, Singer, & Willett, 1990). The time to modify instruments is during pilot work, not data collection.

Principle Number 4: Collect Data Prospectively Whenever Possible

Retrospective data collection is fraught with problems. Even simple information collected by retrospection—on the occurrence and spacing of critical events—can be unreliable. Although important one-time events—such as college graduation and taking a first job—may be remembered indefinitely and highly salient events—such as a leave of absence from teaching—may be remembered for several years, habitual events—such as daily work activities—are forgotten almost immediately (Bradburn, 1983). The longer the interval, the greater the errors. Respondents forget events entirely (*memory failure*), they remember events as having occurred more recently (*telescoping*), and they drop fractions and report even numbers or numbers ending in 0 and 5 (*rounding*). Memory failures lead to underreporting, telescoping to overreporting, and rounding to both. When measurement depends upon respondents' judgment—as when measuring mental "states" like efficacy and satisfaction—gathering high-quality retrospective data is next to impossible.

Data should be collected retrospectively only when this does not challenge their reliability and validity. Much can be learned about better ways to collect retrospective data by examining methods for doing so developed in other disciplines (e.g., Freedman, Thornton, Camburn, Alwin, & Young-DeMarco, 1988; Means, Swan, Jobe, & Esposito, 1991). These studies and others show that better data are obtained when questions about when an event occurred are linked to contextual questions about *where* and *why* it happened (Bradburn, Rips, & Shevell, 1987; Featherman, 1980; Friedman, 1993). Narrative formats—as opposed to a series of individual questions—are especially useful when combined with cues designed to aid recall.

Principle Number 5: Collect Data Beginning in Multiple Base Years

Recognizing that the teaching force is changing over time, NCES fields the SASS and TFS every three to four years. NCES must retain this commitment in the new study. Rather than following a single cohort of teachers from a single base year, we recommend that it follow several cohorts from multiple base years (Janson, 1981; Wall & Williams, 1970).

Including multiple base years prevents analysis from being stymied by the "Age-Period-Cohort" problem (Mason & Fienberg, 1985). A teacher's place "in time" is marked by (a) an entry cohort (the first year of teaching),

(b) the year in the career (first year, second year, known as "age" or "experience"), and (c) the chronological year being described (2000, 2001, known as "period"). All three markers are important. Teachers' responses may reflect the effects of the year they entered teaching (cohort effect), the year of their career (age effect), and the year being described (period effect).

Knowledge of any two dimensions defines the third. Data on the third year of the career for a teacher who entered in 1996 describe the chronological year 1999. This dependence makes it difficult for data analysts to determine unequivocally whether observed differences across teachers should be attributed to cohort effects, year-of-teaching effects, or chronological-time effects. For instance, Mark and Anderson (1985) concluded that teachers hired in the mid-to-late 1970s were less committed to teaching than were previously hired teachers. In a reanalysis of their data, Singer and Willett (1988) showed that this supposed "entry effect" was more likely a "period effect"—large numbers of teachers hired in the later cohorts were laid off, making it only *seem* as though they had lower commitment.

Cross-sectional studies confound all three sources of information about time. Longitudinal studies begun in a single year confound two. It is only by collecting truly longitudinal data in multiple base years that researchers can take steps toward unraveling these effects. Comparing longitudinal data on teachers who enter the study in the base year of data collection with those who enter in a later cohort, for example, allows researchers to isolate the effects of chronological time. We therefore recommend that the new study sample teachers in two (preferably three) base years.

Principle Number 6: Collect Data at All Relevant Levels of the Organizational Hierarchy

Teacher data must be viewed against the backdrop of the environment within which teachers work. Some of these data must come from respondents other than the teachers themselves. Multi-level data allow researchers to explore other respondents' opinions about the school environment and the sampled teachers. Each teacher's principal can provide contextual data on the teacher's position in the school. So, too, data describing the students that the teachers serve can be obtained by including items on the teacher questionnaire (asking the teacher to describe the students) or by building linkages between the new study and ongoing NCES data collection efforts that include students (e.g., NAEP). Other possibilities include in-class observations in selected areas and the use of "work diaries" in which respondents record their time use and interactions over several weeks (Silberstein & Scott, 1991).

Realistically, NCES will be able to create full multi-level data for only a subset of teachers. But by targeting this data collection, decreased generalizability can be repaid by better measurement and tighter focus. Better to have high-quality data on important constructs in a targeted subsample than poor-quality data on the entire group.

Specifying the Target Population: Whom Should Be Studied?

To identify a target population, we must first consider *where* the nation's teachers are in their careers, a complex task because of the heterogeneity of the teaching force and the diversity of teachers' career paths. We begin this section by diagramming the teaching career and then using data from the SASS and TFS to comment on what we can expect to find when this longitudinal study is fielded. We conclude by presenting four distinct possible target populations from which a sample could be drawn.

Diagramming the Teaching Career

For simplicity, think of the teaching career as having annual transition points and, while we may ultimately wish to distinguish teachers who transfer schools from those who stay in one place, also limit the employment options to one of two possible states—"in teaching" or "out of teaching." All teachers begin in the "in teaching" state. Each subsequent year, a teacher may move to either the "in" or "out" state. Over time, the number of possible career paths increases exponentially. After 10 years, for example, each member of an entering cohort could have followed one of $2^9 = 512$ distinct paths!

Regardless of the number of years considered and the resultant myriad paths, all teachers' paths can be classified into one of four mutually exclusive (and exhaustive) categories: Path A, teachers who remain uninterruptedly; Path B, teachers who leave and never return; Path C, teachers who leave and return (and may leave and return again) and

who are currently teaching when observed; and Path D, teachers who leave, return, and leave again (and may return and leave again) and who are *not* currently teaching when observed. Specification of these four categories allows us to summarize the career trajectories of *all* teachers teaching—current and former—in any given year.

Figure 1 uses parallel subtables to present a sample of these paths. Year T, the last column in each subtable, designates the base year of data collection. For simplicity, the figure includes only three entry cohorts (labeled "Early," "Middle," and "Late"). The years of the career between Years 1, 2, and the present (Year T) are accounted

Early-Entry Cohort

Year 1: Year of Hire	Year 2	Year 3	Intervening Years	Year T: Base Year of Survey
In teaching	In	In	As many years as necessary for teachers in this cohort to get to the base year of data collection for the longitudinal study	A: In
				B: Out
		Out		C: In
				B: Out
	Out	In		C: In
				D: Out
		Out		C: In
				B: Out

Middle-Entry Cohort:

Year 1 Year of Hire	Year 2	Year 3	Intervening Years	Year T Base Year of Survey
In teaching	In	In	As many years as necessary for teachers in this cohort to get to the base year of data collection for the longitudinal study	A: In
				B: Out
		Out		C: In
				B: Out
	Out	In		C: In
				D: Out
		Out		C: In
				B: Out

Late-Entry Cohort:

Year 1 Year of Hire	Year 2	Year 3	Intervening Years	Year T Base Year of Survey
In teaching	In	In	As many years as necessary for teachers in this cohort to get to the base year of data collection for the longitudinal study	A: In
				B: Out
		Out		C: In
				B: Out
	Out	In		C: In
				D: Out
		Out		C: In
				B: Out

FIGURE 1. *Understanding who is included in, and who is excluded from, the stock of teachers present in the base year of a survey. Contributions from early-entry (top panel), middle-entry (middle panel), and late-entry (bottom panel) cohorts.*

for in the column labeled "Intervening Years," which represents teachers' occupational states in Years 4, 5, and so on. The full complement of tables necessary for describing the careers of all current and former teachers in any given year would be much larger as a separate table would be needed for each possible year of entry.

Figure 1 highlights two dilemmas that arise when attempting to specify a target population. The first concerns the likely bias associated with sampling from the full teaching stock (Lancaster, 1990). This population, composed of only those teachers "in teaching" in that year (those in the shaded cells in the last column of Figure 1), includes two groups: those who never had a career interruption (Path A) and those who had one or more interruptions, but who are currently teaching (Path C). A sample drawn from this stock will not generalize back to any entry cohort because it omits eligible colleagues who began in the same year but who left teaching at some point and are not currently teaching (Paths B and D). Hoem (1985), following Ryder (1965), calls this bias "selection by virtue of survival." If the teaching force were an enclosed system (like a sink with water pouring in at one rate and draining out at another potentially different rate), the effect of this omission (peers following Paths B and D) could be estimated using concomitant data. But because of perturbations in the system arising from three sources of variation—(a) across chronological years in the size of entering cohorts, (b) across chronological years in the size of non-voluntary exiting cohorts, and (c) in the risk of moving schools and leaving teaching as a function of entering cohorts—it is impossible to estimate the size of the bias.

A second problem concerns the diversity of the current teaching stock. Despite being limited to those following Paths A and C, this group is very heterogeneous. Those following Path A come from many different entry cohorts. Those following Path C not only come from many entry cohorts, they also vary with respect to the number and duration of previous spells. Some might argue that this is a problem of analysis, not design. Information on entry cohort and the number and duration of spells could be used to model the heterogeneity. But because teachers are not distributed uniformly with respect to these attributes, allowing them to vary naturally may render the base year sample too unbalanced for detailed analysis. (See, for example, the difficulties experienced by Arnold, Choy, & Bobbitt (1993) in using the TFS to model attrition.) Heterogeneity in entry year and in the number and duration of previous teaching spells must be considered when specifying a target population.

What Do We Know About the Stock of Current Teachers?

To illustrate the potential impact of these complexities, we examined data from the first SASS and TFS describing the number and duration of teaching spells. Although these cross-sectional retrospective data suffer from all the problems identified earlier, we believe that it is better to base design decisions on some information than simply on educated guesses and conventional wisdom.[2]

Breaks in service, although common, are not the norm (Figure 2, bottom panel). About two thirds (65%) of teachers were in their first spell, another quarter (24%) were in a second spell, and only 11% were in a third spell or higher. This distribution, however, differs by initial entry cohort. Less than 15% of teachers hired in the 1980s reported a break in service whereas, among those hired before 1965, more than half reported at least one interruption. Part of this association certainly results from the varying lengths of time available to these cohorts for leaving and returning. But regardless of its cause, the figure documents the heterogeneity in spell number in the current teaching stock. Because about one third of the teaching stock reports at least one break in service, NCES must either collect data describing these teachers' *entire* career histories (both in and out of teaching) or set them aside from the target population. Without knowing about their early years on the job, it is impossible to situate them at an appropriate place on a time axis during analysis.

The top panel of Figure 2 presents the distribution of experience within spells. To facilitate comparisons, the graph is drawn so that within-spell percentages sum to 100. For teachers in the first spell (the first density), median years of experience is 12, with approximately equal quarters in early career (Years 1–5), early mid-career (Years 6–12), later mid-career (Years 13–19), and late career (Year 20 and over). The distribution of spell length for teachers with one or more breaks in service (the other two densities) is similar, but skewed toward shorter lengths. The early years—first through fifth—

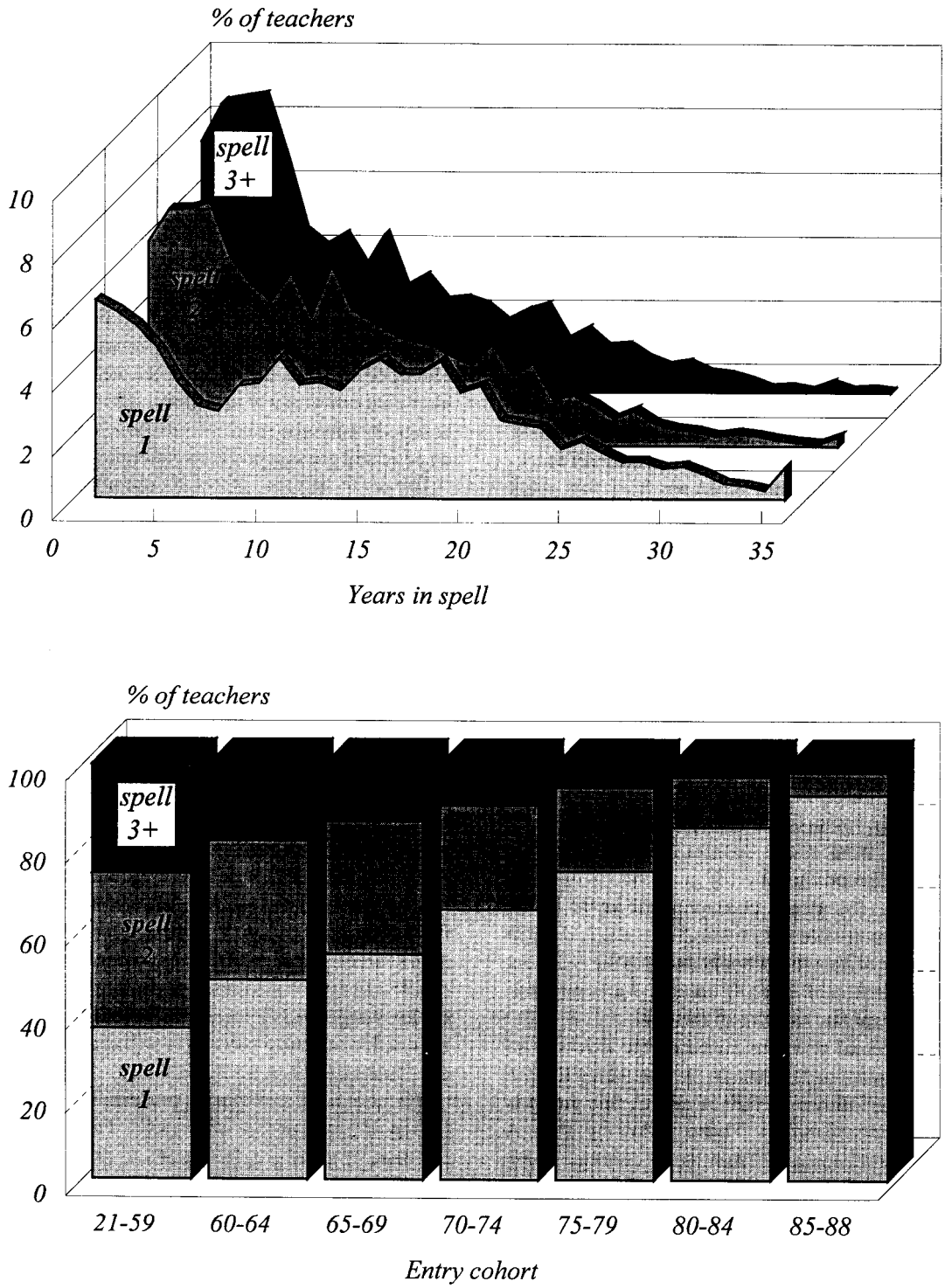

FIGURE 2. *Top panel: distribution of years of teaching experience in the 1987–1988 Schools and Staffing Survey by teaching spell. Bottom panel: distribution of the number of teaching spells in the 1988–1989 Schools and Staffing Survey by entry cohort.*

TABLE 1
Four Possible Samples for a Proposed Longitudinal Study of the Teaching Career

Full stock	1st year	2nd year	3rd year	4th year	5th year	—	nth year
1st spell	✓	✓	✓	✓	✓	✓	✓
2nd spell	✓	✓	✓	✓	✓	✓	✓
—	✓	✓	✓	✓	✓	✓	✓
nth spell	✓	✓	✓	✓	✓	✓	✓
First-year teachers							
1st spell	✓						
2nd spell	✓						
—	✓						
nth spell	✓						
Beginning teachers							
1st spell	✓	✓	✓	✓			
2nd spell	✓	✓	✓	✓			
—	✓	✓	✓	✓			
nth spell	✓	✓	✓	✓			
Stratified stock							
1st spell	✓		✓		✓		✓
2nd spell	✓		✓		✓		✓
—	✓		✓		✓		✓
nth spell	✓		✓		✓		✓

account for one third (35%) of those with only one break and 43% of those with more than one break. Median spell length for these two groups is therefore shorter—nine and seven years, respectively.

Four Possible Sampling Designs

Table 1 presents four possible plans: a stock sample, a sample of first-year teachers, a sample of beginning teachers, and a stratified stock sample. Each represents a valuable resource for studying teachers' careers. The conundrum is that each focuses on a different phase and has its own limitations.

Selecting a *stock sample* is perhaps the most obvious approach. This design's advantages include the ease with which the universe can be listed, the inclusion of teachers with diverse histories, the ease with which subgroups (e.g., beginning teachers, bilingual teachers) can be oversampled, and the potential overlap with a base year SASS. But a stock sample has serious disadvantages. First, by including *all* teachers, the study is unfocused. Oversampling alone (given cost constraints) is unlikely to ameliorate this problem fully. Second, the sample is very heterogeneous with respect to entry cohort, number and duration of career interruptions, and length of time in current spell. Although attempts could be made to reconstruct each teacher's career history from licensure to the base year of data collection, older teachers would be reporting on events long gone. Third, because the teaching stock in any year is disproportionately composed of teachers in the early years of their careers (as shown in Figure 2), senior teachers would be represented in relatively small numbers. Fourth, the "selection by virtue of survival bias" renders generalizability near impossible.

The second design abandons the quest for broad coverage and goes to the other extreme: sample *teachers in the first year of their current spell,* a period known to be critical (Johnson, 1991; Murnane et al., 1991). A longitudinal study of first-year teachers is the "cleanest" of the four options for it describes a well-defined sample over a well-defined career period. The most prominent strength of this design is that the time clock starts at the same moment for all sampled teachers. Notice that this first-year teacher

study does not require everyone to be in the first spell—the sample also includes teachers returning after a career interruption (see Table 1). We include this group because more than half of all newly hired teachers are returnees, and inclusion of them eliminates definitional problems concerning who is really a first-year teacher—those who have never taught or those who have not taught in the past 10 years.

A first-year teacher study has its own shortcomings. It is highly specific—teachers in their first year of a spell compose only 6% of the teaching force. Because any given base year is unlikely to have special policy significance, the study can become less relevant as the entry cohort ages—a criticism that can be leveled at virtually every national longitudinal study ever conducted (including NCES' NLS-72, NELS, and High School and Beyond). This study also excludes colleagues of the returnees who began at the same time as the second- and third-spell teachers but who did not interrupt their careers or who did not return by the base year of data collection. This omission prevents researchers from investigating, for example, whether returning teachers more closely resemble their former colleagues who never left or their newly hired peers. A third shortcoming is that because many years must pass before the sample reaches mid- and late-career, this phase of the career cannot be investigated well.

The two other designs—a beginning-teacher study and a stratified-stock sample—fall between these extremes. In modifying the core designs, of course, each acquires the shortcomings of the other. The beginning-teacher study abandons the specificity of the first-year study by including all beginning teachers—say, in the first five years of their current spell. The stratified stock sample abandons the full coverage of the full stock sample by choosing teachers in specific years of their current spells. In both modifications, heterogeneity with respect to entry cohort makes generalization difficult (other than to that of the base year using the subsample of first-year teachers).

The major advantage of the beginning-teacher sample over the first-year teacher sample is improved coverage: this target population comprises approximately 30% of the teaching force. As with the first-year sample, teachers vary with respect to spell, underscoring the need for gathering previous employment histories for those not in their first spell. The disadvantage of expanding coverage beyond teachers in the first year of service is the bias in sampling by virtue of survival. By focusing on teachers near the beginning of their current spell, however, retrospection is likely to be less fallible. The stratified-stock sample also has broader coverage than the first-year teacher study, but the use of teachers with varying amounts of experience makes it difficult to reconstruct equally precise career histories. Can teachers in their 20th year as adequately reconstruct the time-varying experiences and characteristics of their first year of teaching as those hired just two years ago?

What Is Our Recommendation?

To help decide among the options, we ask: *which aspects of time are most relevant for the phenomena under study?* For a multipurpose study of teachers, this question has no single answer. Developmental studies of adolescents provide a useful contrast. Most studies of school-aged children select students in particular grades and follow them over time. To unravel age, period, and cohort effects, data collection may be replicated in multiple base years. This wisdom of this strategy derives from the near identity between a child's age and grade, the salience of a child's grade/age in affecting outcomes, and the developmental trajectories that children follow. When studying children, birth cohort effects and period effects are assumed to be small in comparison to grade/age effects.

With teachers, however, there are even more aspects of time to contend with, the same aspects of time are not as salient for teachers as they are for children, and the prioritization of the different aspects of time varies across research subdomains. Breaks in service of varying duration expand our conceptualizations of time beyond the classic three parameters (entry year, year of the career, chronological year) to include spell number, length of previous spells, and length of previous out-of-teaching breaks (Willett & Singer, 1995). When exploring job satisfaction, should chronological age be used instead of years of experience? If we use years of experience, should they accumulate across spells or be reset each time a teacher re-enters? Are years of experience really important after 10 or 15 years of teaching? For students, 10th grade differs fundamentally from 11th; for teachers, is there a similar differ-

ence between the 10th and 11th years in the classroom? Huberman (1989) goes so far as to argue that any aspect of time itself may be an empty variable in the study of teachers.

We therefore conclude that a stratified-stock sample—the design of choice for longitudinal studies of children—is not optimal for this study. Neither is the selection of a simple-stock sample from the entire pool of teachers teaching in a specific base year a good alternative because of selection bias.

We believe instead that the two studies of beginning teachers—those in their first year or those in their first few years—are preferable. The early years are clearly critical, and for many researchers likely to analyze these data, the changes of greatest interest are most likely to occur during this time. From a methodological perspective, the study of first-year teachers is superior because it eliminates selection bias. From a policy perspective, however, the desire to investigate mid-career teachers provides a strong incentive to expand the target population to begin with teachers in the first few years of service. The potential for bias persists, but if data from one of the already-fielded SASSs or TFSs are used to select the teachers, the bias can be estimated.

The Time Dimension: How Often, and For How Long, Must Data Be Collected?

For the new study to support modern methods for longitudinal data analysis, it needs to measure change and event occurrence well. We begin this section by discussing how many waves are needed to measure change well by considering (a) the *shape* of the individual growth trajectory, (b) the *precision* with which we want to measure change, and (c) the *reliability* with which we want to distinguish teachers based on these changes. We then discuss requirements for studying event occurrence. We conclude with recommendations concerning data requirements across time.

The Shape of the Individual Growth Trajectory

In the simplest case, growth is assumed to be linear over time and the *individual growth model* representing each teacher's data contains two *individual growth parameters*—an *intercept* and a *slope*—representing his or her initial value and rate of change. Heterogeneity in change across teachers is reflected in inter-individual variation in growth parameters. In linear growth, for example, between-person heterogeneity in change is reflected in variation across teachers in intercepts and slopes.

Conceptually, the specified growth model can be viewed as a *within-person regression model* representing individual change over time. Regardless of the method of estimation, an individual growth model is fit to each teacher's empirical growth record and, as in any regression analysis, to fit the model, we need at least one more data point than there are unknown parameters in the model (Willett, 1994). A linear growth model requires at least three waves. More complex models increase the requirements. Quadratic models need at least four waves; cubic models, five. Different constructs will likely require different growth models. Change may be linear in one domain and exponential in another. Fewer waves will be required in domains where growth is less complex. To determine the number of waves, then, we must use the literature, expert testimony, or, better yet, pilot data to make educated guesses about the shape of the growth trajectories expected for each variable under study.

The Precision of Estimates of Growth

These minimal data requirements leave one degree of freedom per person for estimating model goodness-of-fit (including residual sums-of-squares and standard errors). But just because a model can be fit does not mean that its parameters are estimated well. The precision of the model's estimates improves with additional waves (Cook & Ware, 1983). We illustrate this relationship in the top panel of Figure 3 where we display the standard error of the individual rate of change (in residual standard deviation units) as a function of the number of waves of data collected.[3] The more waves, the smaller the standard error of the linear slope, reflecting improved precision in measuring the rate of change. For example, as the number of waves increases from 3 to 5 (all else being equal), the standard error of the slope is more than halved. As it increases further to 7 or 8, the standard error is quartered.

The Reliability With Which Change Can Be Measured

Above, we focus on intra-individual (Level 1) change. Yet we are also interested in how change differs across people, known as *inter-individual*

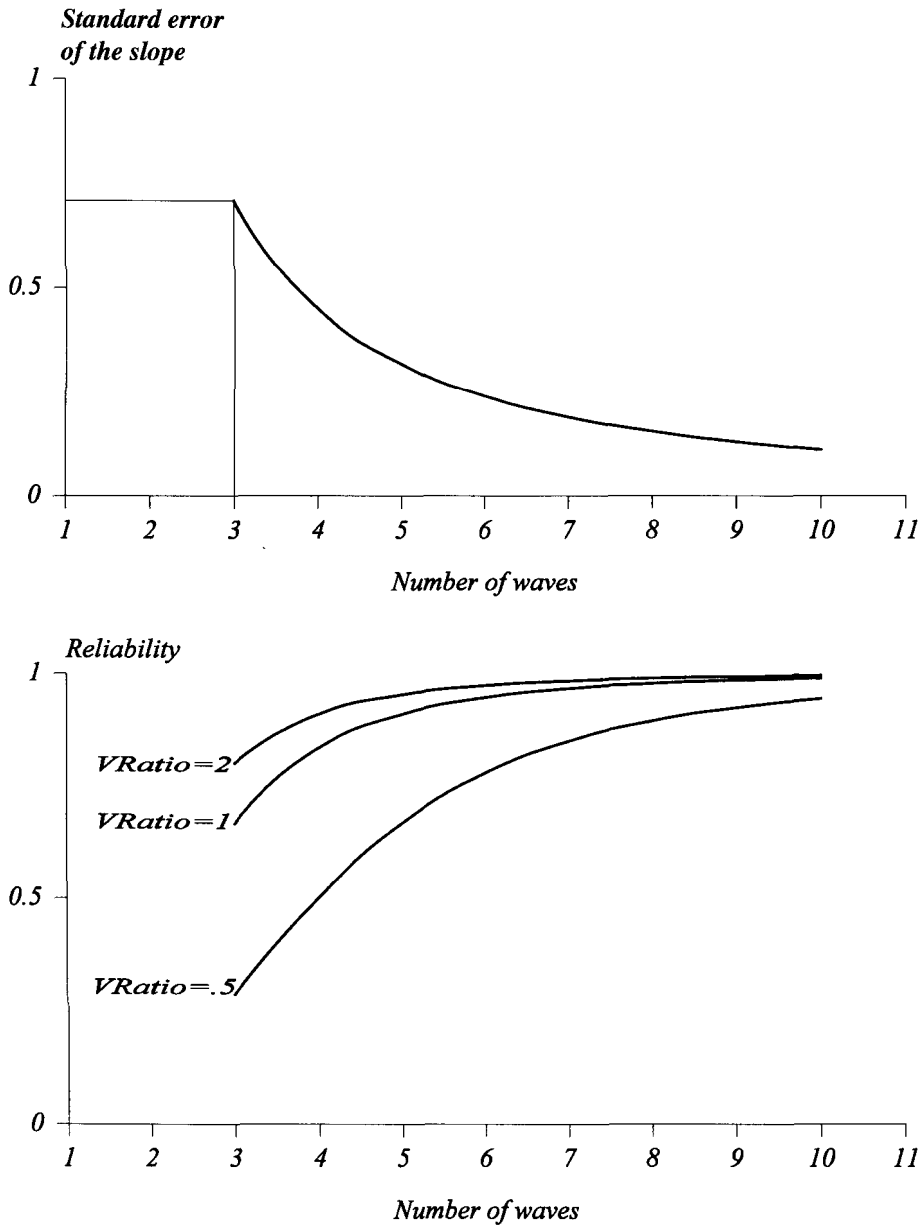

FIGURE 3. *The precision and reliability with which individual change can be measured, displayed as a function of the number of waves of data collected. Top panel: standard error of individual rate of change (in units of residual standard deviation). Bottom panel: reliability of rate of change. Assuming linear individual growth, ordinary least-squares estimation of the rate of change and equally spaced waves of longitudinal data.*

(Level 2) *differences in change* (Bryk & Raudenbush, 1992; Rogosa & Willett, 1985). At Level 2, we ask how individual change is related to predictors, and we are usually concerned with the *reliability* with which change can be measured because reliability describes the extent to which people can be distinguished from each other based on their individual changes (see Rogosa et al., 1982; Willett, 1988).[4]

Increasing the number of data collection waves increases the reliability with which we can detect inter-individual differences in change. Because reliability is a group-level parameter, however, it also can be affected by the hetero-

geneity in true change in the population. If there is no heterogeneity in true change (everyone is growing at the same exact rate), the reliability of change is zero (because we cannot distinguish people on the basis of their changes). As heterogeneity in true change increases, so does reliability. The conceptual disadvantage of reliability is that it confounds the effect of Level 1 measurement-error variance with Level 2 heterogeneity in true change. When measurement-error variance is large or heterogeneity in true change is small, reliability will be low; when either measurement-error variance is small or heterogeneity in true change is large, reliability will be high.

In the second panel of Figure 3, we display the reliability of the rate of change against the number of waves of data for three values of the ratio of the population variance of true linear slope to measurement-error variance, *VRatio*.[5] In any study, *VRatio* can take on widely disparate values for different constructs. With imprecise measures, there may be little inter-individual heterogeneity in true change. In this case, measurement-error variance will be high, and the variance of true change will be small, leading to a value of *VRatio* that is much less than 1. The bottom curve in the second panel represents this situation, with *VRatio* set to .5, a value not uncommon in practice (see Williamson, Appelbaum, & Epanchin, 1991). In this case, when three waves of data have been collected, the reliability of change is very low (0.28).

When measurement is more precise (or when heterogeneity in true change is large), *VRatio* will be larger than 1. The top curve in the second panel represents this situation, with *VRatio* set to 2. In this case, three waves of data yield a reasonably reliable measure of change (0.80). The middle curve represents a *VRatio* of 1, a value that corresponds roughly to the longitudinal measurement of teacher satisfaction, measures of which we constructed by matching four items from the 1987–1988 SASS with corresponding items on the 1988–1989 TFS. In this case, three waves of data provide a measure of change that is moderately reliable (0.67).

Inspection of the lower panel of Figure 3 leads to three design conclusions. First, increasing the number of waves increases the reliability with which change is measured. For a measure like teacher satisfaction, for example, which we estimate to have a *VRatio* of 1, adding two waves to the basic three increases the reliability of change from 0.67 to 0.91. Second, the impact of additional waves is greater for designs that begin with fewer waves. Once again, when *VRatio* = 1, increasing the number of waves from 3 to 5 increases the reliability of change by 38% (0.67 to 0.91); a further increase to 7 waves increases reliability by only 5% (0.91 to 0.97). Third, the impact of adding waves is disproportionately felt when *VRatio* is small—in other words, when measurement-error variance is high or heterogeneity in true change is small, adding a few extra waves will dramatically improve reliability. For a *VRatio* of 0.5, for example, an increase in the number of waves from 3 to 5 yields a 133% increase in reliability (from 0.28 to 0.67); for a *VRatio* of 2, in contrast, the increase is only 19%.

How Long Should Teachers Be Observed?

After dismissing the answer "forever" as impractical, we ask: what is the *minimum study duration?* One way to address this question is by examining the rate of transition in the teaching force. Lacking national longitudinal data (which this new study will provide), we have pieced together pseudo-longitudinal summaries using the SASS and TFS. As before, these admittedly imperfect summaries (based here on two waves of data) can be useful for study planning.

Figure 4 presents estimated discrete-time pseudo-hazard functions describing the risk that a teacher will (a) leave teaching and (b) move schools in each year of the first, second, and third (or higher) spells in teaching. These plots are not true discrete-time hazard functions: they do not describe the risks teachers experience as they move from one year to the next; rather, they display the risk of leaving teaching (and of moving schools) for teachers at each level of experience in the current spell.[6]

Inspection of the profiles suggests that the career of first-spell teachers divides into three epochs of roughly equal duration: the early (0–12), middle (12–24), and later (25–36) years. First-epoch risks tend to be elevated initially, but decline with time. During this period, teachers are more likely to move schools than leave teaching entirely. As the years pass, the risks of moving and leaving converge so that by Year 12, both events are equally likely. Risk equality is approximately preserved during the middle epoch until the risks of leaving increase with the onset

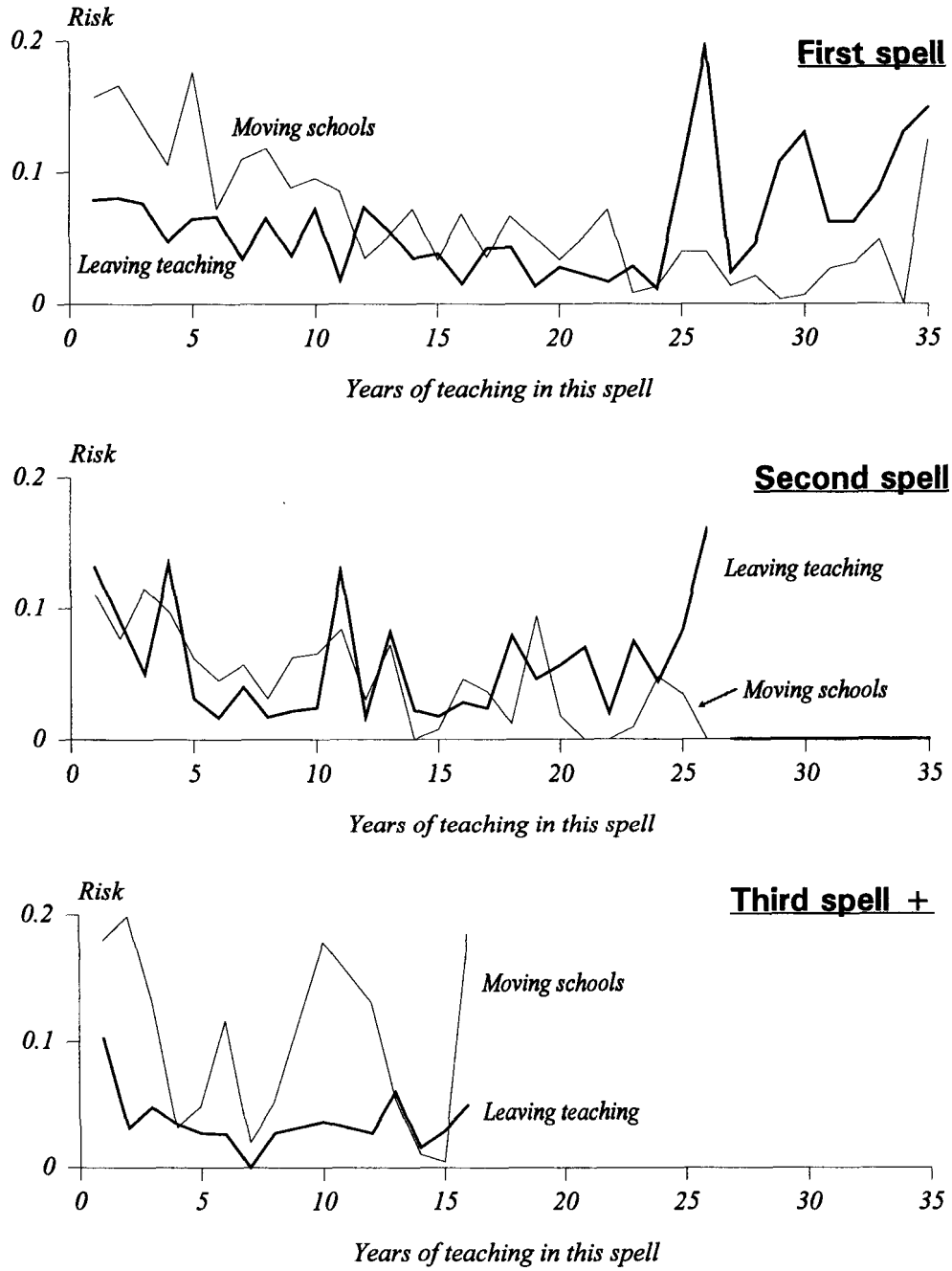

FIGURE 4. *Pseudo-longitudinal discrete-time hazard functions displaying the estimated risk of moving schools, and leaving teaching, for each year of experience by spell. Based on two-wave data from the 1987–1988 Schools and Staffing Survey and the 1988–1989 Teacher Follow-Up Survey.*

of retirement in the third epoch. In the second spell, a U-shaped profile is also found but epochs are less distinct and, with smaller sample sizes, sampling variation more apparent. The small number of teachers in advanced spells prevents us from reaching firm conclusions about this group. As a whole, however, most of the action in all three panels occurs early on, suggesting that to be useful to researchers studying career transitions, the new study must gather data on the early years, reinforcing our recommendation of defining beginning teachers as the target

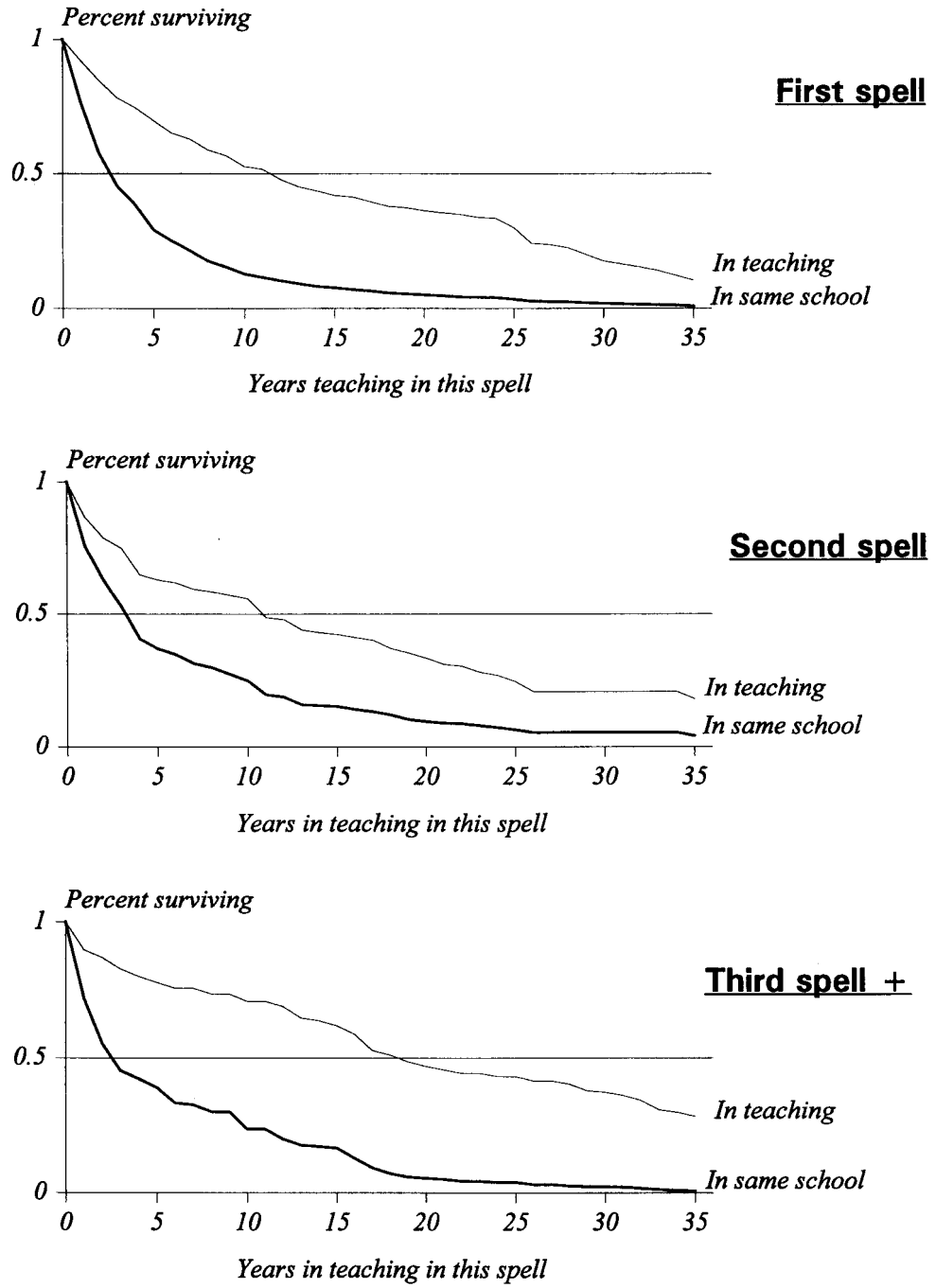

FIGURE 5. *Pseudo-longitudinal discrete-time survival functions displaying the probability of staying in teaching, and staying in the same school, longer than each labeled year of experience by spell. Based on two-wave data from the 1987–1988 Schools and Staffing Survey and the 1988–1989 Teacher Follow-Up Survey.*

population. Finally, the obvious prevalence of "moving" over "leaving" (particularly in the early years) reinforces our recommendation that the new study track teachers across schools.

Figure 5 presents discrete-time pseudo-survival functions associated with the risk profiles in Figure 4. In each panel, notice that the probability teachers will stay in teaching is consistently higher than the probability that they will stay in

the same school—that is, in every year of the career, teachers are more likely to survive in *any* school than in *any particular* school. Also apparent are the dramatic changes in survivorship that occur during the first 10 to 12 years. The horizontal line drawn at a survival probability of 50% helps estimate the time when half of the population of teachers is likely to have experienced each event (estimated median lifetimes). On average, teachers in each spell switch schools quite readily—after about three years on the job. They take much longer to quit teaching entirely, however. Median lifetimes for teachers in their first and second spells are approximately 11 years.

What Is Our Recommendation?

We can use these median lifetimes and statistical power analysis to estimate an appropriate study duration. Singer and Willett (1991, Table 1, p. 277) show that the minimum sample size required to achieve reasonable statistical power varies considerably as a function of study duration.[7] When follow-up extends to the median lifetime, the same power can be reached with half the sample size that would be needed if follow-up extended only to half the median lifetime. If follow-up is twice the median lifetime, the same power can be achieved with one third the sample size. The authors recommend that sample members be followed for at least the median lifetime—and preferably longer.

We therefore recommend that cohorts of beginning teachers in all spells be followed for at least 12 years. This provides decent statistical power at reasonable sample size and ensures that about half of the teachers will leave teaching during observation, creating a pool of returnees available for study. And because the average teacher switches schools approximately every 3 years, a 12-year window provides much data on movement across jurisdictions.

During this 12-year measurement window, we further recommend collecting six equally spaced waves of data. Six equally spaced bi-annual observations on each teacher ensures that growth of limited curvilinearity can be modeled well and that change can be measured with reasonable precision and reliability. Within this overall plan, NCES is encouraged to consider additional brief (inexpensive) annual contacts to ensure that essential data on the intervening years can be obtained reliably.

Conclusion

In this paper, we have outlined major methodological issues relevant to the design of a large-scale longitudinal study of teachers' careers. We have tried to be broad—describing general goals and design principles—and specific—providing our judgment of the best of the possible alternatives. We hope this presentation will generate a debate on these issues, thereby informing the planning process.

The teaching profession, and education in general, is in transition. Changes in teacher education programs, alternative pathways, mentoring programs, and new certification requirements are being adopted throughout the country. Add to this the many administrative changes—school-based decision-making, school choice, charter schools—and it is clear that a longitudinal study has to be flexible to keep abreast of the times. One danger of longitudinal research is that, as the sample ages, findings become less relevant to current practice. For this reason, we have argued for re-initiating the proposed study in at least two, and preferably three, base years. But there are other dangers. If we do not collect data on policy changes, unexplainable variation will escalate. Without knowing that a teacher mentoring program has been established, for example, how can we explain improved collegiality? Rather than treat innovative policy implementation as a nuisance, we must regard these initiatives as natural experiments and systematically embed evaluations within the larger study (Fienberg & Tanur, 1987). True evaluation of alternative programs may not be possible, as random assignment of teachers (and students) to program seems unlikely, but much may be learned by describing program effects and tracking their natural course of development. At the very least, we recommend that data be collected from teachers, principals, and districts on the implementation of alternative programs and that the sample be augmented by subsamples drawn from districts in which innovative programs are fielded, perhaps in collaboration with state and school district themselves.

We conclude with a plea for what is probably the most important issue for NCES to consider: *the need for extensive pilot studies conducted well in advance of data collection.* We lack too much data right now to flesh out all the study's details. In this paper, we have tried (insofar as it

is possible) to use cross-sectional and two-wave data sources to make guesses about the data that might be collected. Many questions remain. How changeable are teachers' reports of their attitudes towards their jobs? Are attitudes more variable than behavioral indicators of how teachers spend their work time? Can teacher quality be measured? Answers to these questions require a substantial planning phase with targeted empirical data collection. Although pilot studies can be expensive, we prefer to think of them as early intervention. Better to conduct extensive pilots before data collection than to discover that an elegant strategy fails when confronted with the real world of schools and fieldwork.

Twelve years of bi-annual data on a sample of beginning teachers should produce an impressive database with the potential to yield much new knowledge about the teaching profession. Unlike studies based on administrative records, it collects data from teachers themselves. Unlike analyses of subsamples of teachers included in other national surveys, it gathers data from a range of informants so that teachers' work lives can be understood in context. And unlike the current SASS and TFS, it is truly longitudinal—not relying on retrospective recollection and limited one-year prospective reports. We now turn to our readers, hoping that discussion of these topics among researchers, policymakers, and government officials leads to better decisions concerning options.

Notes

The order of the authors was determined by randomization. We thank Dan Kaspryzk and Sharon Bobbitt of the National Center for Education Statistics and four anonymous reviewers for their suggestions and insight. An earlier version of this paper was presented at the annual meeting of the American Educational Research Association, New Orleans, LA, in April 1994.

[1] We use "change" here broadly, to encompass changes in diverse domains such as educational status, family status, employment status, attitudes, knowledge, and behavior.

[2] We used SASS-I and TFS-I because, at the time of writing, the TFS for SASS-II was unavailable. Although point estimates may differ between the surveys, we expect conclusions to be similar.

[3] Figure 3 assumes linear individual change, equally spaced measurement occasions, and ordinary least-squares estimation of individual rate of change. Letting T be the number of waves of data collected, the standard error of the slope (in units of residual standard deviation) is

$$\frac{s.e.\,(slope)}{s.d.\,(residual)} = \sqrt{\frac{12}{T(T^2-1)}}$$

[4] The reliability of change is defined as the proportion of the variance in observed change that is variance in true change in the population (see Rogosa et al., 1982; Willett, 1988).

[5] Under the assumptions listed in text, the population reliability of the ordinary least-squares estimate of the linear rate of change is given by

$$\rho(\hat{\pi}) = \frac{(\sigma_\pi^2 / \sigma_\epsilon^2)}{(\sigma_\pi^2 / \sigma_\epsilon^2) + \dfrac{12}{T(T^2-1)}}$$

where σ_π^2 is the population variance of the true linear growth rate, σ_ϵ^2 is the population measurement-error variance, and T is the number of waves of data collected (see Willett, 1988).

[6] These profiles were constructed piecemeal from pairs of two-wave data points in the 1987–1988 SASS and 1988–1989 TFS. For instance, we compared the SASS and TFS responses of teachers in the first year of the first spell and estimated the probability that they left teaching between the years. These estimated probabilities were then plotted against years of teaching in each spell.

[7] Singer and Willett (1991) assume effect sizes of magnitudes typical in the study of the teacher career (that is, 1.5 or 2; see Murnane et al., 1991), a two-group comparison at the 0.05 level, a two-tailed test, power of 0.80, and a uniform population hazard function.

References

Arnold, C. L., Choy, S. P., & Bobbitt, S. A. (1993). *Modeling teacher supply and demand, with commentary.* Washington, DC: National Center for Education Statistics.

Billingsley, B. S. (1992). *Teacher retention/attrition: Issues for research.* Washington, DC: National Center for Education Statistics.

Billingsley, B. S. (1993). Teacher retention and attrition in special and general education: A critical review of the literature. *The Journal of Special Education, 27*(2), 137–174.

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

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

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.* Newbury Park, CA: Sage.

Choy, S. P., Bobbit, S. A., Henke, R. R., Medrich, E. A., Horn, L. J., & Lieberman, J. (1993). *America's teachers: Profile of a profession.* Washington, DC: National Center for Education Statistics.

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

Duncan, G. J., & Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review, 55,* 97–117.

Featherman, D. L. (1980). Retrospective longitudinal research: Methodological considerations. *Journal of Economics and Business, 32,* 152–169.

Fienberg, S. E., & Tanur, J. M. (1987). Experimental and sampling structures: Parallels diverging and meeting. *International Statistical Review, 55,* 75–96.

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

Friedman, W. J. (1993). Memory of time of past events. *Psychological Bulletin, 113,* 44–66.

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

Goodson, I. F., & Hargreaves, A. (Eds.). (1996). *Teachers' professional lives.* Washington, DC: Falmer Press.

Grissmer, D. W., & Kirby, S. N. (1992a). *Designing the teacher follow-up survey: Issues and content.* Washington, DC: National Center for Education Statistics Press.

Grissmer, D. W., & Kirby, S. N. (1992b). *Patterns of attrition among Indiana teachers: 1965–1987.* Santa Monica, CA: Rand Corporation.

Grossman, P. L. (1990). *The making of a teacher: Teacher knowledge and teacher education.* New York: Teachers College Press.

Hafner, A., & Owings, J. (1991). *Careers in teaching: Following members of the high school class of 1972 in and out of teaching.* Washington, DC: National Center for Education Statistics.

Hanushek, E. A., & Pace, R. R. (1995). Who choose to teach (and why)? *Economics of Education Review, 14*(2), 101–117.

Heyns, B. (1988). Educational defectors: A first look at teacher attrition in the NLS-72. *Educational Researcher, 17*(3), 24–32.

Hoem, J. M. (1985). Weighting, misclassification, and other issues in the analysis of survey samples of life histories. In J. J. Heckman & B. Singer (Eds.), *Longitudinal analysis of labor market data* (pp. 249–293). New York: Cambridge University Press.

Huberman, M. (1989). *The lives of teachers.* New York: Teachers College Press.

Ingersoll, R. M., Han, M., & Bobbitt, S. (1995). *Teacher supply, teacher qualifications, and teacher turnover: 1990–91.* Washington, DC: National Center for Education Statistics.

Janson, C. G. (1981). Some problems of longitudinal research in the social sciences. In F. Schulsinger, S. A. Mednick, & J. Knop (Eds.), *Longitudinal research: Methods and uses in behavioral science* (pp. 19–55). Boston: Martinus Nijhoff Publishing.

Johnson, S. M. (1991). *Teachers at work.* New York: Basic Books.

Kennedy, M. M. (1992). The problem of improving teacher quality while balancing supply and demand. In E. E. Boe & D. M. Gilford (Eds.), *Teacher supply, demand, and quality: Policy issues, models, and data bases* (pp. 65–108). Washington, DC: National Academy of Sciences Press.

Lancaster, T. (1990). *The econometric analysis of transition data* (Econometric Society Monographs). New York: Cambridge University Press.

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

Little, J. W., & McLaughlin, M. W. (Eds.). (1993). *Teacher's work: Individuals, colleagues, and contexts.* New York: Teachers College Press.

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

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

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

McLaughlin, M. W., & Oberman, I. (Eds.). (1996). *Teacher learning: New policies, new practices.* New York: Teachers College Press.

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.

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

Murnane, R. J., Singer, J. D., Willett, J. B., Kemple, J. J., & Olsen, R. J. (1991). *Who will teach?: Policies that matter.* Cambridge, MA: Harvard University Press.

National Academy of Sciences. (1992). *Teacher supply, demand and quality: Policy issues, models, and data bases.* Washington, DC: National Academy Press.

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

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

Ryder, N. B. (1965). The cohort as a concept in the study of social change. *American Sociological Review, 30,* 843–861.

Shulman, L. S. (1992). Directions for the future. In E. E. Boe & D. M. Gilford (Eds.), *Teacher supply, demand, and quality: Policy issues, models, and data bases* (pp. 287–290). Washington, DC: National Academy of Sciences Press.

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

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

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

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

Theobold, N. (1990). An examination of the influence of personal, professional, and school district characteristics on public school teacher retention. *Economics of Education Review, 9,* 241–250.

Theobold, N. D., & Gritz, R. M. (1992). *Understanding the supply of elementary and secondary teachers: The role of the Schools and Staffing Survey and the Teacher Follow-Up Survey.* Washington, DC: National Center for Education Statistics.

Wall, W. D., & Williams, H. L. (1970). *Longitudinal studies and the social sciences.* London: Heinemann.

Weiss, I. R. (1992). *Reflections on an SASS longitudinal study.* Washington, DC: National Center for Education Statistics.

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Rothkopf (Ed.), *Review of research in education* (Vol. 15, pp. 345–422). Washington, DC: American Educational Research Association.

Willett, J. B. (1994). Measurement of change. In T. Husen & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed., pp. 671–678). Oxford, England: Elsevier Science Press.

Willett, J. B., & Singer, J. D. (1989). Two types of question about time: Methodological issues in the analysis of teacher career path data. *International Journal of Educational Research, 13,* 421–437.

Willett, J. B., & Singer, J. D. (1991). From whether to when: New methods for studying student dropout and teacher attrition. *Review of Educational Research, 61,* 407–450.

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

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

Williamson, G. L., Appelbaum, M., & Epanchin, A. (1991). Longitudinal analyses of academic achievement. *Journal of Educational Measurement, 28,* 61–76.

Authors

JUDITH D. SINGER is a professor at Harvard University Graduate School of Education, Larsen Hall, Appian Way, Cambridge, MA 02138. She specializes in quantitative methods, research design, and longitudinal analysis.

JOHN B. WILLETT is also a professor at Harvard University Graduate School of Education, Gutman Library, Appian Way, Cambridge, MA 02138. His specialties are quantitative methods, longitudinal analysis, and research design.

Received July 18, 1995
Revision received July 11, 1996
Accepted July 19, 1996