Design and Implementation Issues in Longitudinal Research

Larry H. Ludlow
Joseph Pedulla
Emile Mitescu Reagan
Sarah Enterline
Mac Cannady
Stephanie Chappe

Boston College

Abstract: To meet demands for accountability, most schools and departments of education at institutions of higher education (IHEs) gather information on their current students and graduates. This paper describes issues to consider when designing a longitudinal data collection and management system, drawing on seven years' experience developing such a system. The recommendations provided stem from an attempt to use data collected for accountability more broadly to look at the specific issue of teacher retention and attrition. Recommendations include: involve stakeholders in all phases of the project; minimize staff turnover; attend fastidiously to

record keeping and documentation; minimize changes in measures; expect that costs will be under-estimated; and gather data on multiple cohorts.
Keywords: longitudinal research; surveys; teacher education; accountability

Diseño y cuestiones de implementación en investigaciones longitudinales
Resumen: Para satisfacer las demandas de rendición de cuentas (accountability), la mayoría de las escuelas y departamentos de educación en instituciones de educación superior educación (IHE) reúnen información sobre sus estudiantes actuales y sus egresados. En este trabajo se describen problemas a tener en cuenta para diseñar sistemas de gestión y modelos para recolectar datos longitudinales aprovechando la experiencia de siete años el desarrollo de dichos sistemas. Las recomendaciones proporcionadas surgen de un intento de utilizar datos recogidos para la rendición de cuentas de manera más amplia para examinar la cuestión específica del retención y deserción de docentes. Las recomendaciones incluyen: involucrar a las partes interesadas en todas las fases del proyecto, minimizar la perdida de personal, ser meticulosos en el mantenimiento de registros y documentación, minimizar los cambios en las medidas, anticipar que los costos serán subestimados, y recopilar datos sobre múltiples cohortes.
Palabras clave: investigación longitudinal; encuestas; formación docente; rendición de cuentas

Design e problemas na implementação de pesquisas longitudinais
Resumo: Para atender às demandas de prestação de contas (accountability), a maioria das escolas e departamentos de educação nas instituições de ensino superior (IES) reúne informações sobre seus alunos atuais e graduados. Este artigo descreve problemas a serem considerados na elaboração de sistemas de gestão e modelos para a coleta de dados longitudinais, utilizando a experiência de sete anos no desenvolvimento de tais sistemas. As recomendações decorrem de uma tentativa de usar os dados coletados de um programa de prestação de contas mais amplo para analisar a questão específica da retenção e de desgaste dos professores. As recomendações incluem: envolvimento das partes interessadas em todas as fases do projeto; minimização da perda de pessoal; precisão e cuidado na manutenção de registros e documentação; redução das alterações nas medidas; previsão de que os custos poderão ser subestimados; e recolhimento de dados sobre coortes múltiplas.
Palavras-chave: pesquisa longitudinal; pesquisa de opinião; formação de professores, prestação de contas

Introduction

In response to the demand for accountability, most schools and departments of education in institutions of higher education (IHEs) are gathering information on their current students and graduates (Wineburg, 2006). Teacher education programs face particular criticism and demands to demonstrate their effectiveness. Are IHE teacher education graduates better prepared than teachers who come from other pathways into the profession? Do the pupils of IHE-prepared teachers perform better than the pupils of teachers who have entered the profession from other pathways? What proportion of IHE teacher education graduates enters the profession? What proportion who enter stay in the profession? What differentiates those who stay in teaching from those who leave?

While IHEs are collecting data to address these questions, Wineburg (2006) concluded that most "do not appear to be able to organize and interpret the data in ways that would provide an effective response to outside mandates" (p. 7). Specifically, Wineburg (2006) found that "lack of data management systems, lack of access to data, and lack of a consistent methodology to gather and analyze data were often cited as impediments" (p.7). Furthermore, although Schalock, Schalock & Ayres (2006) commend the American Educational Research Association (AERA) Panel on Research

and Teacher Education (Cochran-Smith & Zeichner, 2005) for having "crafted numerous recommendations for overcoming the limitations repeatedly found in the studies reviewed", they also note that the AERA Panel's recommendations "do not carry the detail that provides guidelines to a researcher engaged in the actual design and implementation of a study that is intended to reflect the recommendations advanced" (p. 103-104). To address these, and other questions related to IHE accountability, longitudinal studies are needed.

The purpose of this article is to offer advice on how to capitalize on the strengths of a longitudinal design, while anticipating and coping with the problems. Specifically, this article presents design and implementation details based on seven years' worth of experience developing and conducting a longitudinal teacher education tracking and data management system to address program accountability and, subsequently, teacher retention issues (Ludlow, Mitescu, Pedulla, Cochran-Smith, Cannady, Enterline, & Chappe, 2010a; Ludlow, Pedulla, Cannady, Mitescu, Chappe, Enterline, Loftus, Cantor & McMahon, 2010b; Mitescu, Ludlow, Pedulla, Cochran-Smith, Cannady, Enterline, Chappe, Cantor, Hu, J. & Loftus, 2009). We share our experiences and offer recommendations for conducting longitudinal educational research, in general, and research on teacher education, in particular, to meet accountability demands.

Regardless of the original purpose of any study conducted by an IHE (e.g. self-study or external program accountability), we expect that the school or department of education will be interested in capitalizing on these initial efforts at accountability studies by extending its work to address other research questions. We believe that much of what we discuss here can be applied to other IHEs as they examine their own data management processes and systems with a view toward getting the most information from them.

A search of the literature for advice on how to conduct "longitudinal research" quickly reveals four general classes of references: a) general advice on, among other things, the merits of longitudinal versus cross-sectional designs, the need to reduce attrition, how to build data files and select statistical software, and ways to report the results (e.g. Cook & Ware, 1983; Menard, 1991; Bronner, 2002); b) technical advice on building scales appropriate for developmental studies, "best practice" procedures for handling missing data, new developments and extensions in statistical models, working examples, and considerations in selecting between alternative models (e.g. Little, Preacher, Selig & Card, 2007; Little, Bovaird & Card, 2007; Grimm, Pianta & Konold, 2009); c) calls for additional longitudinal research within a specific existing research area (e.g. Singer & Willett, 1996; Vanderstay, 2006; Avey, Luthans & Mhatre, 2008); and d) specific project applications offering detailed practical advice on addressing linguistic challenges and different value systems among the research team, how to create consent forms and how to format surveys, when to schedule mailings, how to maintain data security while addressing Institutional Review Board requirements, reviewing the 'worthiness' of new items, retaining participants, and maintaining the funding for the project (e.g. Bynner, 2005; Byles, Dobson, Bryson & Brown, 2007; Loxton & Young, 2007). The present paper falls into category (d).

In this article, we discuss the characteristics of longitudinal studies, including their relative strengths and weaknesses. Next, we present the history and context of our project, share our experiences, and discuss practical issues of conducting longitudinal research. Specifically, we address the importance of developing a conceptual framework and issues concerning project staff, stake-holder involvement, population and sampling, instrumentation, data management and analysis, and cost. We conclude with key recommendations for conducting longitudinal research. It is our hope that through our experience (including our mistakes), our hindsight can be your foresight.

Characteristics of Longitudinal Research

We follow White and Arzi's (2005) definition of longitudinal design: "A longitudinal study is one in which two or more measures or observations of a comparable form are made of the same individuals or entities over a period of at least one year" (p. 138). In longitudinal research, Bauer (2004) differentiates between "cohort" and "panel" studies. Specifically, in cohort studies, which examine data across a set of individuals who share common experiences at the same time (e.g., the same graduating class), data are aggregated "across all members of the cohort, thus excluding the opportunity to include individual variance" (p. 76). On the other hand, in panel studies, "the same individuals are studied at two or more points in time. Panel studies thus allow the researcher to account for individual change" (p. 76). In addition, Bauer (2004) notes, "longitudinal panel designs often include multiple cohorts and thus offer the most robust set of analyses. Multiple cohort panel designs enable analysis of age, period, and cohort effects; description of developmental and historical change; analysis of temporal order events; and causal analysis" (p. 76).

As noted above, the major strength of longitudinal designs is the ability to look at change over time in the measures on those being studied. Shadish, Cook, and Campbell (2002) suggest that these "longitudinal designs allow examination of how effects change over time … and are frequently more powerful than designs with fewer observations" (p. 267). Since measures are obtained on the same people at multiple time points, one can look at correlates of change, patterns of change, and other elements of interest as they relate to change over time.

However, longitudinal designs have their own special problems (e.g., time, resources, data management, attrition, and cost) and are often difficult to implement (White & Arzi, 2005). By their very nature they extend over a period of time with data being gathered throughout this period. In addition, longitudinal designs make it essential to gather data from as many people in the population as possible at each time point so that change over time can be examined with confidence. This increases the cost and resources needed and complicates data management and analysis.

With the preceding as a backdrop, we present next the context in which our multiple cohort panel design study took place. We provide the history of the project together with a description of the cohorts, the panels, and the data collected.

History of the Project

The Boston College Lynch School of Education (BC LSOE) Ford Research Team began in 2008 as a continuation of the research team originally created as part of a Carnegie Teachers for a New Era (TNE) project that ran from 2003 to 2008. Shortly after its selection as a TNE site, the BC LSOE group formed a multi-disciplinary Evidence Team responsible for developing instruments and conducting research to assess the success of the program and to foster evidence-gathering activities. One of the team's first tasks was the specification of a conceptual framework (Cochran-Smith and the Boston College Evidence Team, 2009) to guide its work. This framework informed, among other things, a portfolio of mixed-methods studies, including a series of BC LSOE surveys examining teacher candidates'/graduates' perceptions, experiences, beliefs, and practices over time (Ludlow, Pedulla, Enterline, Cochran-Smith, Loftus, Salomon-Fernandez, and Mitescu, 2008).

Taken as a group, the BC LSOE surveys have a number of purposes, including: (1) surveying teacher candidates' (and later teacher graduates' and beginning teachers') perceptions, expectations and beliefs about teaching and their expected career trajectories in the profession; (2) surveying candidates'/graduates' sense of preparedness for teaching and their evaluations of the usefulness of various aspects of the preparation program at different points in time; (3) surveying beginning teachers' practices and strategies once in the classroom and their self-reported assessments of the

impact of various BC TNE program elements on their own learning, their teaching practices, and the learning of their pupils; and, (4) establishing a data base that would link the responses of each teacher candidate to other evidence-gathering activities and data bases (e.g. observational data concerning teaching practices, retention data). Each survey serves a specific purpose that is related to the time at which it is administered. Beginning in the spring of 2004 and continuing through the fall of 2010, there have been a total of 28 administrations of the Entry, Exit, One-Year Out (YO), Two-Year Out (TYO), and Three-Year Out (THYO) Surveys. See Figure 1 for the survey administration schedule.

Year	Survey				
2004	Exit				
2005	Exit	OneYear Out			Entry
2006	Exit	OneYear Out	TwoYear Out		Entry
2007	Exit	OneYear Out	TwoYear Out	ThreeYear Out	Entry
2008	Exit	OneYear Out	TwoYear Out	ThreeYear Out	Entry
2009	Exit	OneYear Out	TwoYear Out	ThreeYear Out	Entry
2010	Exit	OneYear Out	TwoYear Out	ThreeYear Out	Entry

Figure 1. Survey Administration Schedule: Accountability and Retention Studies

These surveys provide the data for the heart of the longitudinal quantitative analyses conducted to date. Specifically, cross-sectional comparisons of successive entering cohorts demonstrate the consistency of the entering candidate's characteristics. Cross-sectional comparisons of successive graduating cohorts reveal improvements in various aspects of the teacher education program (Ludlow, et al., 2008). In addition, panel analyses following each graduating cohort reveal perceptions of program experiences subsequently influenced by classroom practice, and replications of procedures and analyses across cohorts serve to confirm and strengthen cohort experiences and findings (Ludlow et al, 2010b).

Recently we became interested in using our existing TNE-developed data management system to study teacher retention. Since half of all teachers in urban settings are estimated to leave the profession in the first five years (Ingersoll, 2003), IHEs have an obligation to provide schools and pupils with a steady source of highly qualified teachers. Furthermore, since institutions spend extraordinary amounts of time, energy and resources training teachers, they have a responsibility to determine if they are doing a good job. Part of that determination involves an examination of who stays and who leaves teaching and why. Accordingly, the data gathered primarily for the purpose of evaluating the teacher preparation program are now being used to investigate factors underlying our graduates' teacher retention and attrition circumstances (Ludlow et al, 2010b).

Practical Issues and Considerations in Longitudinal Research

In this section, we present practical considerations that need to be addressed in any longitudinal

study. Specifically, one must have a clear conceptual basis for the research, ensure that the staffing needs of the project are adequately addressed, involve all parties potentially affected by the results, clearly define the population(s) of interest, give due consideration to how the sampling frame(s) relates to the population(s), have clarity about the primary purposes for the instrumentation, have a plan for ongoing data management integrated with the data analysis plan, and estimate the cost of all these factors in the event that adjustments need to be made to stay within budget.

Conceptual Framework

It is critical to create a conceptual framework that defines the purpose of the longitudinal research within a well-established literature (see, for example, Cochran-Smith and the Boston College Evidence Team, 2009). The purpose of this framework is to guide the initial specification and inevitable refinement of the important questions to be addressed. Such a framework will, consequently, highlight the fact that educational research is a multi-disciplinary effort requiring flexible research designs capable of gathering and analyzing data (i.e., a portfolio of studies) that directly address the nature of the questions and are not dependent on the specific tendencies and training of the investigators. Once the framework is established, then it is possible to specify the demographics of the population, the measurement instruments to construct, the observations to gather, the kinds of qualitative and quantitative analyses to perform, and the significance and implications of the study.

Furthermore, the conceptual framework provides continuity in plans and objectives across the span of the project. This point is important because over the course of a longitudinal project faculty may take sabbatical leaves, different jobs, new administrative responsibilities, or even retire. Similarly, staff and graduate students may come and go. These situations can fracture the cohesion and commitment of the remaining group members, and this can lead to confusion and disruption unless there is a clear framework linking future studies with those that have already occurred.

Project Staff

Generally, the less the amount of staff turnover, the smoother the project will operate. However, as noted above, project staff will enter and leave for different reasons. Furthermore, team members will have different levels of investment, whether professional or personal, in the success of the project. Reducing turnover and strengthening commitment can be accomplished through the creation of professional situations tailored to benefit each member of the project. For example, providing research assignments for administrative staff seeking new opportunities and responsibilities, ensuring that faculty have relatively quick opportunities for publications and conference presentations, and providing resources for graduate student development through workshop training can strengthen long-term commitment and intellectual investment in the project. In addition, graduate students brought in with diverse skill sets should be provided opportunities to utilize those skills in creative and independent ways, expected to work with and learn from others with different backgrounds, and mentored and advanced into more senior levels of responsibility over the course of their time on the project.

Another factor to consider is the experience of the senior personnel when they become members of the team. Although bringing in relatively inexperienced graduate students may work well over the course of a three to four year involvement, a specialist such as a database developer/manager should be hired based on experience—not on an individual's potential to develop the skills needed for the work. Furthermore, when considering the qualifications of the data manager, think broadly about the individual's educational research background and fit with the specific research agenda. In our experience, it helps when the data manager is familiar with both the teacher education department's mission and routine statistical concepts and procedures.

This concern over broad experience is relevant, too, when considering the qualifications of the statistician/quantitative data analyst. An analyst unfamiliar with teacher education interests and objectives cannot simply be told: "run the data and tell us what's important." This means that the project's staffing interest should not necessarily be primarily about finding a person with advanced statistical training. Rather, it may be that the project needs someone who can listen carefully, translate the substantive interests of teacher educators into actionable and relevant quantitative analyses, and can then communicate technical procedures and results clearly and meaningfully.

Stakeholder "Buy-in"

A longitudinal program accountability and/or retention study will span years from the initial proposal, early development of the research team and its assessments, the gathering of data for a given cohort followed by replications across subsequent cohorts, to the final analysis and report writing stage. The simple initial enthusiasm and passion of the team will not suffice to ensure the success of a longitudinal project, where success is defined in terms of generating evidence that feeds back and enhances the strengths of the program and its graduates.

All stakeholders (e.g., university, school and department administrators; program faculty and students; and funders) must understand the painstakingly slow and tedious nature of longitudinal research. It isn't simply that "change takes time"—it takes time to: define what is important to measure, develop and revise measures, gather data and conduct the analyses, determine what should/can be changed within a curriculum/program of studies, devise and implement those changes, and analyze and evaluate post-change data. The "buy-in" of all stakeholders is essential for maintaining the commitment and resources needed to maximize the project's likelihood of success.

This buy-in can be addressed in various creative ways. Meetings can be arranged with student organizations to review early drafts of assessments and lunch-time forums can provide ongoing awareness of the project and opportunities to discuss findings. Program faculty can be informed early-on about the project, solicited for their expectations about the project, and included in the initial development and review of assessments. Later, they can provide a critical audience for discussion of the results and the implications for curriculum or program changes.

Faculty can integrate the project into their curriculum in various ways that continuously highlight its professional relevance and significance (e.g., higher education policy courses), and its operational details (e.g., classroom observation protocols discussed in qualitative research courses, scale development used as examples in psychometrics courses, longitudinal data analysis incorporated into statistics courses). Administrators, too, can be kept informed over the years by regular updates about conference presentations, publications, invited addresses, inquiries about collaborations, project related dissertations, and suggestions for admission recruitment materials. It is not sufficient to assume that stakeholders outside the immediate research team know the significance and contributions of the project to the quality of the program; without continued updates, their participation and support may wane.

Population and Sampling

It may seem that defining your population of interest is an obvious initial task that should be quite straightforward. We encountered the following issues when defining our population.

Census or sample. How many people will be studied? This age-old research question can easily be handled by tracking every candidate and graduate. Granted, taking a census at each time point may be a resource drain. It will, however, reduce the noise of critics who challenge the representativeness of voluntary respondents. It will also start the study off with a larger pool of participants who will still be available after multiple years of participation.

A simple voluntary sample of convenience, although cheaper and quicker, is likely to produce fewer participants than projected and their representative nature will be suspect. The results will be open to challenge by reviewers and stakeholders. A well-designed random sample may be feasible, but it introduces its own unique analytic problems with sampling weights and design effects that must be handled properly. This design will also not address the smaller sample size issue that is a particular concern with longitudinal designs.

Cohorts or "tracks". Being aware of the graduation paths and having a clear idea of who will be studied will make the data collection and analysis processes run more smoothly. It will also ensure that students are administered the appropriate surveys at the proper time. Not all teacher education candidates who begin their program of studies in the same semester will graduate at the same time. Not all candidates become teachers. How will the candidates and graduates be grouped and tracked: by time of matriculation; graduation (which date if a candidate has more than one degree from your institution); or expected graduation (i.e. combine summer graduates with prior spring graduates or the following year's graduates)?

For example, one of our first discoveries in conducting the follow-up retention study was that previously defined TNE cohorts, or groups of individuals operationally defined by their year of graduation and Exit survey administration, no longer made sense. This became evident when longitudinal files were created and some individuals were found with multiple surveys of the same type (e.g., two Exit surveys), since many students receive more than one degree from LSOE. As a consequence, we created a "track" variable so that we had only one set of responses that corresponded to the most complete survey data over the Exit, YO, TYO, and THYO administrations. With this variable we pull data based on both the person and a specific sequence of surveys and can be sure that we do not count the individual in the study more than once, as happened when we analyzed "cohorts."

Attrition. No consideration of longitudinal studies is complete without some discussion about how to reduce attrition of subjects. The inevitable fact is that some participants will leave the study. The standard efforts to keep people responding include multiple emails, regular postal mailings, and phone calls—for us, this sometimes meant as many as ten attempts to reach someone. Contact efforts by phone (sometimes coordinated as "phone banks") were followed by cover letters, at various times, from the Dean, department chair, project principal investigators, and the head of the practicum office. We also informed the incoming classes about the project, and the importance of their participation in it, in their required first semester Professional Development Seminar; provided materials through various courses throughout their academic tenure; submitted materials to their student newsletters; reminded them through Practicum Office flyers; and provided them with updates through the Mentoring and Induction Office. All these interactions stressed the significance of the project to the school and the profession, and the critical need for each person to respond and participate for at least the three immediate years after graduation. Striving for census-level returns across 28 survey administrations, we attained, through this intense effort, response rates in excess of 95% for the Entry and Exit Surveys, 65% for the One- and Two-Year Out Surveys, and 60% for the Three-Year Out Surveys.

The concerted efforts to apprise teacher candidates and graduates of the importance of the study and the subsequent logistics involved in maintaining contact with them had a profound positive impact on the response rates. These were expensive activities, but they reduced numerous threats to the internal and external validity of the project.

Complete or partial longitudinal records. With longitudinal data one has to decide whether to look only at individuals with complete data across all the years or to study all individuals with any data at any particular time. Studying persons with complete data will result in smaller n's for the analyses, but the characteristics of the people stay the same at each time point. Including everyone available

for analysis at any given time point will generate larger n's but will create concerns over the comparability of the sample at different times.

Making the decision about which of these two ways of analyzing the data you will employ early on, if it is reasonable to do so, may influence the data collection process. For program accountability purposes, it may not matter if the sample is somewhat different at each time point after graduation. But if the study is to address teacher retention, then it is critical to follow-up and track each individual graduate. Each person's story about why they stayed or left teaching becomes more important the longer the study continues.

Operationally defining participants. Similar to defining the population, identifying who would be called a "teacher" seemed obvious, but our experience doing so was much more complicated than expected. For example, an early measurement challenge was to simplify, yet adequately cover, the multitude of education-related employment options that could be interpreted, or not, by people as teaching. Various formal definitions of *teacher* provide specific perspectives. For example, the Oxford English Dictionary suggests that a teacher is "one who or that which teaches or instructs; an instructor; one whose function is to give instruction, especially in a school." Asking people if they meet the dictionary definition may seem reasonable, until you ask them to list their job title or describe what they do. At this point you encounter problems trying to designate who is or is not a teacher.

Once we decided on what constituted a "teacher", we discovered that one's teaching status at a given time point often conflicted with what was reported on a person's multiple surveys and "contact" forms. Contact forms were typically sent to graduates in January with the surveys following later in the year. These two sources can provide differing information regarding what a respondent says about their teaching. For example, one respondent indicated on the contact form that she was a teacher but on the survey elaborated that she was a private tutor for two months—a description that we did not accept as "teaching." Multiple situations like this necessitated development of a "Conflict Identification and Resolution Rules" document (Ludlow et al, 2010b).

Decide exactly how a teacher is to be defined based on their type of job (e.g., full-time teacher, permanent substitute, building substitute, etc.), their response patterns on the surveys (e.g., did they skip the section for "teachers only"), and other sources used to collect teaching status information about them (e.g., the contact forms). Once these definitions and decision rules are set, avoid changing them if at all possible. The more consistent the definitions or rules are over time, the less chance there will be of misclassifying individuals and having to re-run analyses.

Once "teaching" is defined, the type of teaching patterns to study have to be specified. That is, using Ingersoll's (2003) classifications, is a person a "stayer", "leaver" or "mover"? Initially we thought there would be reasonably sized groups of graduates who had: (a) not taught at all; (b) taught one year and then left teaching ("leaver"); (c) taught two years and then left teaching ("leaver"); (d) taught three years at the same school ("stayer"); and, (e) taught three years at some combination of different schools ("mover"). Fortunately, we had the system capability to determine how many people were in each of these categories and how many people were missing information on any of the other permutations of stay/leave/move. Unfortunately, for statistical purposes, we discovered we had low numbers of teachers who left teaching once they started and even lower numbers who stayed in teaching for three years but changed schools.

Studying only "stayers" and "leavers" offers analytical simplicity but it may miss nuances important in teachers' career trajectories. For example, the "stayers" category is still only defined for the finite number of years being examined by us; this may lead to results that do not capture completely the differences between "stayers" and "leavers" had the time frame been longer. Furthermore, "movers" is an increasingly studied career trajectory and may therefore warrant focused data on not only whether teachers move, but also, what type of move it was (e.g., a change

in grade level, content area, public/private, urban/suburban setting) and the reason(s) why teachers decided to move to a different school. Also, will attrition be treated as a one-time event or will teachers be studied who transition in and out of teaching over the course of the project? For example, a person who taught year one, did not teach year two but came back and taught year three was treated by us as a "leaver" after year one because we had so few cases like this—your decision about this case might differ.

Instrumentation

The BC LSOE surveys have excellent psychometric characteristics that are well documented and have been used by numerous IHEs both nationally and internationally (Ludlow et al., 2008). Each of the surveys underwent extensive pilot testing before operational use. Students and graduates tend to score high on the scales and the results provide evidence in support of a strong LSOE teacher preparation program.

One issue that must be considered in the instrument development process is the extent to which the instrument will be constructed to provide feedback on program characteristics that the faculty and administration feel are strengths versus addressing areas known to be weak or missing that the faculty and administration seek to remedy. For example, in the current climate of higher education accountability, in general, and teacher education, in particular, it is surprising that formal coursework in classroom assessment and quantitative data collection and analysis is not a standard teacher education program requirement—even in schools of education with strong quantitative departments. A school in this situation provides an ideal opportunity to gather evidence from teacher candidates and graduates about the quality of the assessment and data analysis training they received. Such a focus might reveal a lack of training that has, among other things, left graduates unprepared to explain standardized test results to parents, utilize online diagnostic assessment tools in their classes, or discuss and plan strategies to meet federal and state Adequate Yearly Progress goals. Such questions on a survey might not receive high "preparation" or "satisfaction" scores initially but those scores would serve as a strong foundation for measuring progress over the coming years should the program address this aspect of its curriculum.

Utilization of Available Variables. Surveys originally designed for one purpose may lack information sought for a subsequent purpose. For example, the BC LSOE surveys were designed for accountability or accreditation purposes but they lack some of the information identified in the literature as useful for studying teacher retention. This situation can lead to issues of omitted variable bias in statistical models used for examining teacher turnover. This potential bias and its attendant lack of sensitivity, construct validity, and statistical power is a critical issue to consider at the start of a longitudinal study.

Specifically, in our case, was it likely that graduates would be followed into their teaching careers and studied in terms of the conditions and factors influencing them as they continue to teach or leave teaching? If we had known that the answer was "yes", then questions about classroom resources, administrative support, mentoring and induction opportunities, continuing professional development, preparation level of the pupils, background characteristics of the pupils and the community, and physical safety conditions of the school and community could have been developed, piloted and revised early in the project and incorporated into the surveys. Work is underway to augment the current surveys but numerous cohorts have already passed through the project.

Measuring change. "If you want to measure change, don't change the measure"—a truism in measurement but one that is hard to follow. In our study the same items and scales appear on the same surveys over time, the same analyses are performed on them (as dictated by our "Survey Analysis Manual"), and the results are maintained in commonly-formatted and annotated "Survey

Results Reports." Every item and scale can be tracked over similar surveys, over the course of a single cohort, and across cohorts. When new items or scales are created (e.g., the Job Satisfaction Scale), they are added in their entirety to a survey and do not replace any existing items in a way that changes any existing scales.

The potential negative consequence of this strategy is that items that are highly inter-correlated or that have little response variance are retained, even though they do not add much additional information. On the positive side, this strategy allows for changes in response patterns over time to be determined based on common items and scales that do not change meaning or composition and that do not require new psychometric analyses.

Yearly "upgrades" to the assessments may seem reasonable but they come with a high price. For example, without elaborate equating adjustments, it may be very difficult to claim that positive program changes have occurred if the assessments are not the same from one year to the next. A strong conceptual framework and extensive pilot testing will help minimize the need for item and scale revisions.

Data management and analysis

Data Management. Attention to the types of data that will be gathered, the kinds of analyses to be performed, the extent to which individual data records will be tracked over time, the potential magnitude and complexity of the database, and the skills and experiences of the data manager will all have an effect on the ease and progress of the analytic aspects of the project. In our experience, the BC LSOE Teacher Education System of Assessment (TESA) that was built to gather and analyze multiple years of teacher candidate data for the TNE program accountability project has also operated efficiently and effectively for the purposes of the subsequent teacher retention project. This system, however, did not exist at the start of the original TNE project. When the surveys were first administered in 2004, the data were stored as Excel and SPSS "flat" files. As the administration of the surveys continued, the data quickly proliferated into an unmanageable array of separate and merged files. The confusion and mistakes that began to occur because of similarly named data files, same-named variables, mismatches in participant demographic codes, changes in missing data replacement strategies, and data transformations for "reversed" coding of items for scoring eventually became intolerable.

The operational details of TESA, our current "relational" database system, are documented in Mitescu, et al. (2009) and Ludlow, et al. (2010a). Since TESA is structured around Filemaker Pro v9.03, it is integrated with direct connections to local admissions, enrollment, program, and practicum data to provide contextual information about the participants and their program experiences. Don't assume that it doesn't matter how the data are stored; thinking "it's all in the computer and problems can be fixed later" can lead to serious data management and analysis challenges later in the project.

Intended and operational analysis plans. Statistical analyses designed to test deductions derived from theory are a hallmark of scientific research. Theory building and testing, however, is not a one-time, all-or-nothing affair. Regardless of how "pure" we may wish to present our final models and results, there is a proper role for exploratory ("digging in the muck") analyses.

Theories provide broad ways of understanding teaching and the general circumstances that may influence a person to become a teacher and then stay or leave the profession. Individual scales and items, however, are the crude tools by which we try to operationalize those circumstances. Given the extraordinary lack of consistency in the instruments used by researchers to study teaching, it is unrealistic, if not impossible, to propose a theory-driven hypothesis about every type of statistical relationship that might be tested in one's longitudinal database.

Granted, corrections to control for Type I errors (falsely rejecting the null hypothesis) are common in multiple comparison situations. The problem with too strict adherence to this strategy is the real likelihood of committing Type II errors (falsely retaining the null). This is particularly an issue when sample sizes are small and statistical power is weak. During the exploratory stage of analysis, it may be reasonable to be more concerned about not missing any potentially differentiating scale or item for the eventual model building exercises than about committing Type I errors. Searching for patterns of relationships and group differences with p-values raised as high as .10 may prove fruitful in the development of intermediate-stage hypotheses and final models.

The point, then, is to work towards a strong and defensible final model but to do it in a series of stages that start simply and become increasingly more complex based on combinations of theory-driven and exploratory analyses throughout the course of the project. After all, the point of multiple cohort studies is to test the strength of the findings and inferences drawn on one cohort by examining the extent to which they are replicable on a second or third cohort.

For example, during one of the first Evidence Team meetings in spring 2004, a team member described the conceptual framework captured in Figure 2 (Cochran-Smith et al., 2009) as a structural equation model.
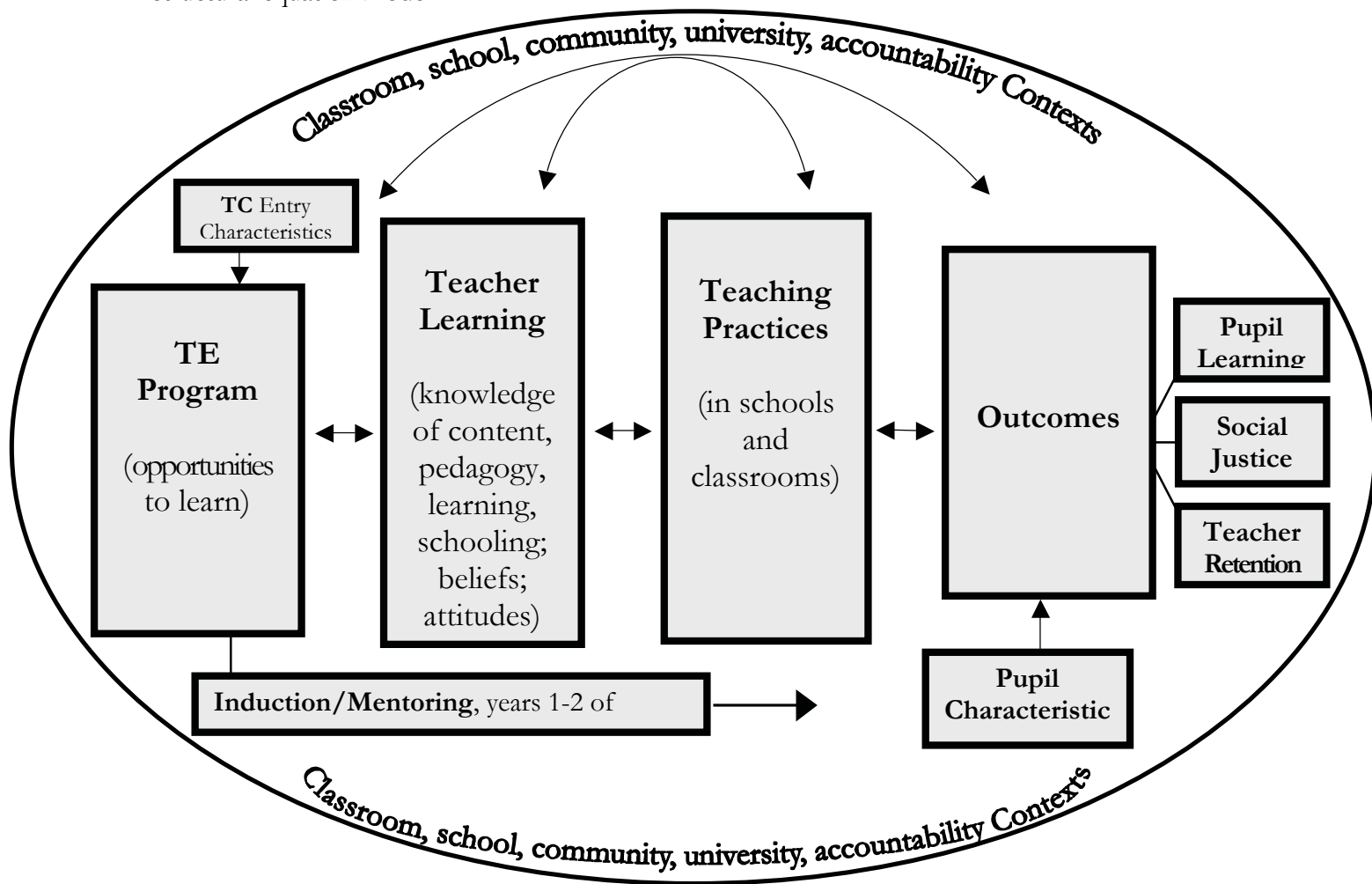
Figure 2. A Conceptual Framework for Assessing Teacher Education

The portfolio of studies presented in Figure 3 of Cochran-Smith et al. (ibid), and expanded upon during the current retention study, may eventually lead to a formal test of the framework as a structural equation model but that model was not the objective of the original study. From the onset of the study (once scale and item-level data from the surveys were available) the analysis plan followed a methodical hierarchical strategy of conducting: a) contingency table chi-square analyses and Pearson correlations testing bivariate relationships; b) independent and dependent means t-tests, and one-way and factorial ANOVAs of group differences; c) simple and multiple ordinary least squares and logistic regressions; and d) survival analyses built on the preceding findings. Many of these hundreds of analyses were driven by theory, many were exploratory, but all contributed under the original conceptual framework to a steady increase in our understanding of simple-to-more complex relationships between teacher candidate characteristics, program experiences, and classroom practice.
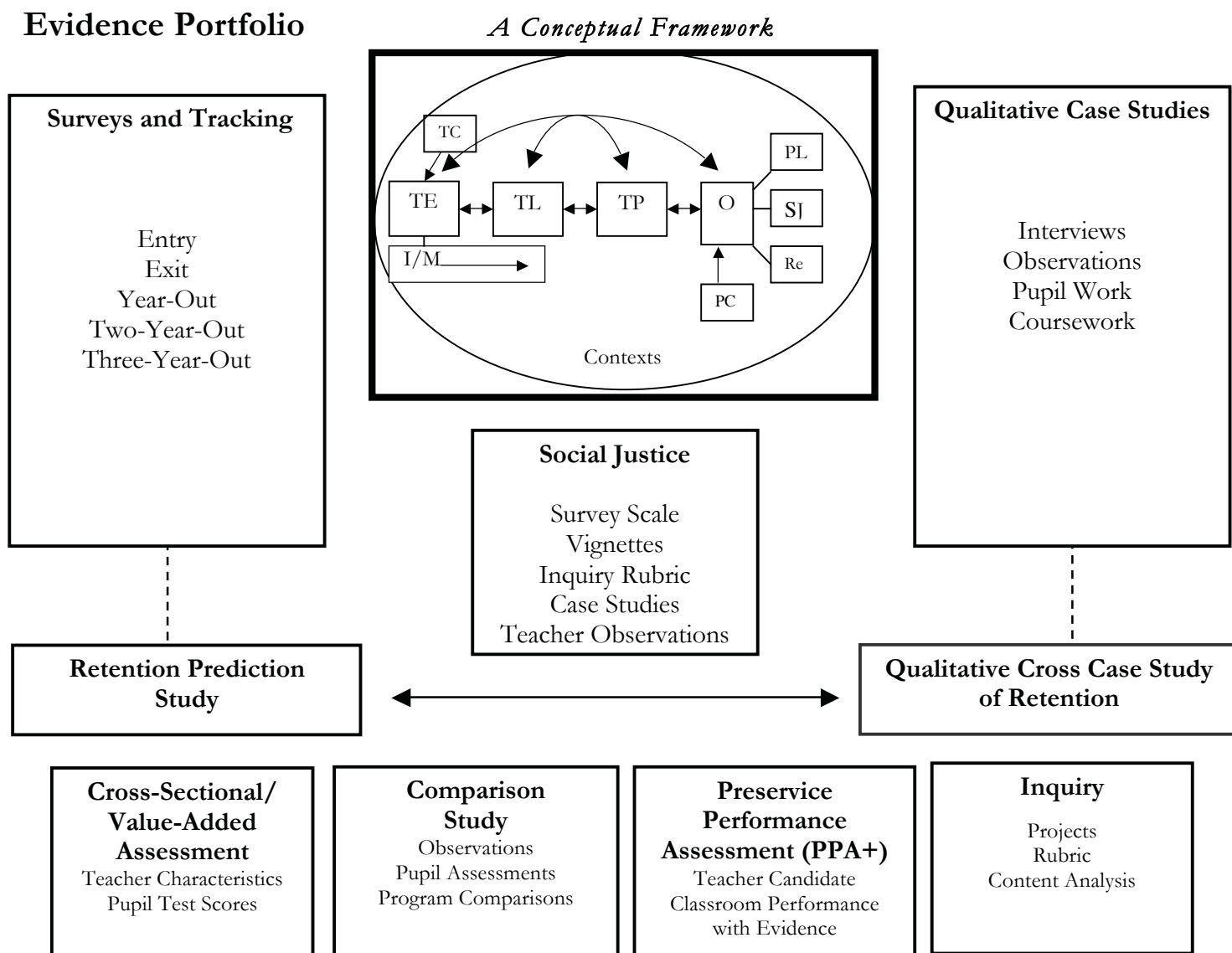
Figure 3. Boston College TNE Evidence Portfolio

We urge avoidance of the "seduction of statistics", i.e. building the most complex statistical model imaginable. Teaching is a complex profession to study, but we recommend starting with simple relationships and interpreting them before adding additional variables to models. As variables are added, see how relationships change, interpret them, and then add more. Regardless of whether the intended final models are structural equation models, hierarchical linear models or event history/survival models, it is always more informative to move from a simple model to a more complex one than it is to start with a complex model and interpret it in isolation or then try to work backwards down to the level of simple relationships.

Record keeping. As noted above, literally hundreds, if not thousands, of statistical tests were conducted (and are still being conducted). Keeping track of, let alone making sense of, all those findings can be a nightmare; this record-keeping issue is often an unforeseen aspect of longitudinal research. Assuming that the project spans years with similar surveys administered multiple times, some form of systematic data analysis strategy for repeated operations has to be created and documented. This will ensure that the same analyses are conducted regardless of personnel changes. Likewise, some form of reporting system has to be created and standardized. This will ensure ready access to similar results regardless of the administration date.

For example, results of analyses of the 28 separate EX, YO, TYO and THYO survey administrations presented in Figure 1 exist as stand-alone bound and online pdf reports for each administration. They are standardized in terms of the analyses performed and format of reporting (as documented in our "Survey Analysis Manual": http://tne.bc.edu).

Furthermore, the longitudinal retention studies require their own set of carefully documented procedures and results. The 2004–2006 tracks were followed for three years after graduation. Each track has comparison groups consisting of those who taught the first year out vs. those who did not teach (compared on the Exit surveys), those who taught the first two consecutive years vs. those who taught the first year and then did not teach the second (compared on the Exit and One-year out surveys), and those who taught the first three consecutive years vs. those who taught the first two years but not the third (compared on the EX, YO and TYO surveys). The three sets of group comparisons on those three different surveys were conducted on each of the '04-'06 tracks. Then because the tracks consisted of small samples, the tracks were aggregated and the analyses re-run. These analyses formed the basis for the prediction of the 2007 track group membership and the final survival analysis models.

Given the number of progressively more complex levels of comparisons conducted in a longitudinal study, it cannot be overstated how essential it is to carefully document procedures and results during each stage of the project. Do not wait until the final stages to try to go back through your records to reconstruct what was done, why it was done, and what the results were. Fully interpret and document what was conducted and found in the preliminary and intermediate stages of the work. Otherwise, details will be lost through poor record keeping and memory lapses.

Cost

Conducting longitudinal research is costly. It would be a mistake to underestimate the costs and jeopardize the quality of the research. Bauer (2004) gives some guidelines:

> Because of the personnel costs, the techniques needed to maintain contact with subjects over time, the costs of incentives, and the need for detailed documentation of data, the time and funding required for longitudinal studies are much greater than those required for cross-sectional designs. A careful analysis of all resources is needed. No matter how advanced the researcher is (that is, has completed many cross-sectional survey studies), the researcher doing his or her first longitudinal study may not know how to budget for unexpected changes in personnel or realize the

time involved in data documentation. A good rule of thumb (passed on anecdotally by colleagues) is to double the time that was originally anticipated for completion of tasks related to the longitudinal study, especially data analysis. Increases in time will also lead to heightened funding needs. (p. 85)

Our experience indicates that Bauer's rule of thumb regarding doubling the original time estimates may still yield an underestimate of the true time involved. For example, the time needed to maintain accurate contact information with all graduates was far more than we ever anticipated. We thought e-mail addresses would work for most graduates; they did not for many. In some cases, we followed bad email addresses with up to ten phone calls (often to parents first to get a current phone number for the graduate). This labor intensive effort was not anticipated. We contained the additional costs associated with this activity by utilizing graduate students and office staff who were already supported and short-term hires of "work study" students. These resources were often provided during different stages of the project by university administration stakeholders who understood the data quality standards of the project.

One must be resourceful in meeting these unanticipated costs. We would caution against eliminating effort to control costs without giving full consideration to the implications of such an action. In the example just cited, had we not put forth the effort to maintain contact with as many graduates as possible, our sample of respondents to the surveys would have been dramatically reduced by the simple fact that we could not get the surveys to large numbers of graduates. This reduction would have had a severe impact on our sample size and therefore, our ability to conduct meaningful statistical analyses. Even if we were a much larger institution that could have obtained a large enough sample for statistical analysis without the extra effort, we would have a potential inherent bias in the sample because of those graduates we could not contact.

Thus, the costs associated with longitudinal research should not be underestimated; they should be weighed against the benefits gained from this research. Many of the important questions and concerns now facing higher education can only be addressed effectively through a longitudinal, multiple cohort design.

Discussion

Lest the reader be left with a sense of hopelessness, let us summarize what we learned from our experience with a view toward optimism. Virtually all IHE's are gathering data to address accountability demands from federal, state and non-governmental accreditation agencies, as well as funders, students, potential students, and parents. These demands often require longitudinal data. Furthermore, the United States Department of Education (DOE) and many state DOEs recognize the strength of having longitudinal databases. For example, although most states do not currently have longitudinal databases linking teachers with their students and containing measures on teachers (e.g. their pathway into teaching) and students (particularly statewide test scores), those that do not are in the process of building them. Once they are available, IHE databases will be able to link with these state databases so that questions about how IHE's graduates' pupils perform can be addressed. We will not get into the many controversies surrounding this issue in this paper; our intent here is simply to highlight the potential for this link between databases.

Longitudinal designs make it essential to gather data from as many people in the population as possible at each time point so that change over time can be examined with confidence. However, longitudinal research comes with its own set of unique benefits and challenges.

In research of this kind, it is essential that the project have a skilled data manager who can build a complex relational database that is capable of tracking people longitudinally. If the data manager has an understanding of teacher education and some basic statistics, the project will benefit

greatly. Furthermore, support and buy-in for the research effort from administrators, faculty, and students is extremely beneficial. In addition, we recommend gathering data from the population, not from a sample, to maximize the number of cases available for longitudinal analysis, since some attrition is inevitable.

We recommend involving program faculty in developing the instruments, identifying research/evaluation questions that are of interest to them, and assisting in the interpretation of the results. All of this provides a spirit of cooperation and interest on the part of the program faculty that enhances the utility of the research. Providing administrators with key findings that they can use with the various audiences that they address regularly (e.g. prospective students, donors, other administrators) gives them data they need to present the program in its best light. Finally, students need to understand the importance of the research and their participation in it. If they understand that they will be followed over time and how the information they provide will help strengthen the program, they are far more likely to continue their participation once they graduate.

We have offered a number of other guidelines or practices in this paper based on what we have learned over the past seven years. These include:

1. To the extent possible, keep the research team intact. Staff turnover is inevitable, but the more you can minimize it, the better.

2. Adhere to the age-old adage: "if you want to measure change, don't change the measure." It is often difficult to stick to this advice, especially when there is evidence that items are not working well or areas of interest have not been addressed. If these situations arise early on in the project, make only the changes deemed to be absolutely necessary, but recognize that you will complicate your analysis and record keeping and that you will lose the ability to measure change on the original variables after the change is made and have the ability to measure change on the new variables only in the future.

3. Use multiple cohorts so that you can replicate findings. Generally, you will be following multiple cohorts as a matter of course anyway. View this as an opportunity to be able to examine the extent to which findings on one cohort (or a group of cohorts) replicate to a newer cohort (or cohorts). In this way, to the extent that the findings replicate, one of the shortcomings of longitudinal research designs, using correlational techniques, will be addressed.

4. Excellent record keeping and documentation are critical. Existing research team members may have memory about how a particular analysis was conducted in previous years or how a scale score was created, but relying only on that memory is likely to lead to errors. Even worse, though, is when a new member of the team must conduct an analysis that replicates one from previous years and has no documentation of what was done previously.

5. The costs associated with longitudinal research are high. We tried to mitigate these costs, which are largely personnel costs, by using existing labor sources (e.g., graduate students who are already supported, clerical staff who can be freed up from other duties) or inexpensive labor sources (e.g., work-study students). We avoided scaling back on efforts to save money, because we judged that we would be giving up valuable information and/or numbers of respondents. The importance of the information and/or respondents outweighed the cost savings in our opinion.

Thus, what may appear, at first, to be a relatively straightforward longitudinal data collection and analysis task is fraught with many potential problems and pitfalls. We hope that by discussing the issues we encountered, presenting the resolutions we utilized to address these issues, and offering recommendations to help avoid or mitigate encountering many of the issues or problems

with longitudinal research, that we will help other IHEs avoid some of these same problems. The area of research into teacher retention is too important; it must be addressed well. Hopefully, lessons we learned will help in ensuring that it is addressed well.

Acknowledgments

References

Adamson, L. & Chojenta, C. (2007). Developing relationships and retaining participants in a longitudinal study. *International Journal of Multiple Research Approaches*, *1*, 137-146.

Avey, J. B., Luthans, F., & Mhatre, K. H. (2008), A call for longitudinal research in positive organizational behavior. *Journal of Organizational Behavior*, *29*, 705–711.

Bauer, K. W. (2004). Conducting Longitudinal Studies. *New Directions for Institutional Research. 121*: 75-90.

Bronner, M. (2002). Conducting longitudinal research. *The Delta Pi Epsilon Journal*, *44*(3), 221-232.

Byles, J.E., Dobson, A., Bryson, L. & Brown, W.J. (2007). Getting started: "Preparing the ground" and "planting the vines" for longitudinal research. *International Journal of Multiple Research Approaches, 1*, 80-91.

Bynner, J. (2005). Challenges for longitudinal research: Synopsis of findings from a scoping study. Retreived from, http://www.longviewuk.com/pdfs/draft_scoping_study.pdf.

Cochran-Smith, M. & Zeichner, K.M. (Eds.). (2005). *Studying teacher education: The report of the AERA Panel on Research and Teacher Education.* Mahwah, NJ: Lawrence Erlbaum.

Cochran-Smith, M. & the Boston College Evidence Team (2009). Reculturing teacher education: Inquiry, evidence, and action. *Journal of Teacher Education. 60*(5): 458-468.

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

Grimm, K.J., Pianta, R.C. & Konold, T. (2009). Longitudinal multitrait-multimethod models for developmental research. *Multivariate Behavioral Research*, *44*, 233-258.

Ingersoll, R. (2003). *Is there really a teacher shortage? A report co-sponsored by the Center for the Study of Teaching and Policy and the Center for Policy Research in Education.* Seattle, WA: University of Washington, Center for the Study of Teaching and Policy.

Little, T.D., Bovaird, J.A., & Card, N.A. (Eds.). (2007). *Modeling Contextual Effects in Longitudinal Studies.* Mahwah, NJ. Lawrence Erlbaum Associates.

Little, T.D., Preacher, K.J., Selig, J.P. & Card, N.A. (2007). New developments in latent variable panel analysis of longitudinal data. *International Journal of Behavioral Development, 31*(4), 357-365.

Loxton, D. & Young, A. (2007). Longitudinal survey development and design. *International Journal of Multiple Research Approaches, 1,* 114-125.

Ludlow, L.H., Mitescu, E., Pedulla, J., Cochran-Smith, M., Cannady, M., Enterline, S. & Chappe, S. (2010a). An accountability model for initial teacher education. *Journal of Education for Teaching. 36* (4) 353-368.

Ludlow, L.H., Pedulla, J., Cannady, M., Mitescu, E., Chappe, S., Enterline, S., Loftus, F., Cantor, D. & McMahon, T. (2010b). Methodological challenges in conducting longitudinal multi-cohort

teacher retention analyses. Paper presented at the American Educational Research Association annual meeting. Denver, CO. April 30, 2010.

Ludlow, L., Pedulla, J., Enterline, S., Cochran-Smith, M., Loftus, F., Salomon-Fernandez, Y. and Mitescu, E. (2008). From students to teachers: Using surveys to build a culture of evidence and inquiry. *European Journal of Teacher Education. 31*(4), 1-19.

Menard, S. (1991). *Longitudinal Research*. Sage University Paper Series on Quantitative Applications for the Social Science, 07-076. Newbury Park, CA: Sage.

Mitescu, E., Ludlow, L.H., Pedulla, J., Cochran-Smith, M., Cannady, M., Enterline, S., Chappe, S., Cantor, D., Hu, J. & Loftus, F. (2009). Building an institution-specific higher education accountability system. Paper presented at the American Educational Research Association annual meeting. San Diego, CA: April 14, 2009.

Schalock, H.D., Schalock, M.D. & Ayres, R. (2006). Scaling up research in teacher education: New demands on theory, measurement, and design. *Journal of Teacher Education. 57*(2), 102- 119.

Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton-Mifflin.

Singer, J.D. & Willett, J.B. (1996). Methodological issues in the design of longitudinal research: Principles and recommendations for a quantitative study of teacher's careers. *Educational Evaluation and Policy Analysis, 18(*4), 265-283.

Vanderstay, S. (2006). Learning from research in criminology and the health sciences. *Reading Research Quarterly, 41*(3), 328-350.

White, R.T. & Arzi, H.J. (2005). Longitudinal studies: Designs, validity, practicality, and value. *Research in Science Education, 35*, 137-149.

Wineburg, M. (2006). Evidence in teacher preparation: Establishing a framework for accountability. *Journal of Teacher Education. 57*(1), 51-64.

About the Authors

Larry H. Ludlow, Joseph Pedulla, Emile Mitescu Reagan, Sarah Enterline, Mac Cannady and Stephanie Chappe
Boston College

Email: "mailto:ludlow@bc.edu"

Larry Ludlow is Professor and Chair of the Department of Educational Research, Measurement and Evaluation. His research interests include the history of statistics, models for understanding course evaluations, estimating teacher education program effects, and Rasch model applications.

Joseph Pedulla is Associate Professor in the Department of Educational Research, Measurement and Evaluation and Senior Research Associate in its Center for the Study of Testing, Evaluation and Educational Policy. His research interests are in evaluating the effectiveness of online delivery of courses, the impact of high stakes testing programs, and examining teacher education program effects.

Emile Mitescu Reagan is a Ph.D. candidate in the Department of Educational Research, Measurement and Evaluation at Boston College. Her dissertation examines teacher candidates' changing beliefs about teaching for social justice as an outcome of teacher education.

Sarah Enterline recently earned her Ph.D. from the Department of Educational Research, Measurement and Evaluation and is currently coordinating the Department of Teacher Education's accreditation process. She also consults regularly with area educational research organizations on topics including research design, survey development and administration, and data analysis.

Mac Cannady is a Ph.D. candidate in the Department of Educational Research, Measurement and Evaluation. He is currently a Senior Research Associate at the Education Development Center and is completing his dissertation comparing the influence of entry attributes and working conditions on early teacher attrition.

Stephanie Chappe is a Ph.D. candidate in the Department of Educational Research, Measurement and Evaluation. Her research interests include examining program effects in higher education.

education policy analysis archives

Please contribute commentaries at http://epaa.info/wordpress/ and send errata notes to Gustavo E. Fischman fischman@asu.edu