

CHAPTER 10**TEACHER LOGS AS A TOOL FOR STUDYING EDUCATIONAL PROCESS***

Brian Rowan
University of Michigan

Eric M. Camburn
University of Wisconsin, Madison

Richard Correnti
University of Michigan

Paper prepared for the conference Using Calendar and Diary Methodologies in Life Events Research, Ann Arbor, Michigan, June 15-17, 2006, sponsored by the National Science Foundation, the National Institute on Aging, and the Survey Research Center at the University of Michigan's Institute for Social Research. To be published in Belli, R., Stafford F. and Alwin, D. (Eds). *Using Calendar and Diary Methods in Life Events Research*. Newbury Park, CA: Sage (under contract).

* The research reported here was supported by grants from the U.S. Department of Education to the Consortium for Policy Research in Education (CPRE) at the University of Pennsylvania (Grant # OERI-R308A60003); the Center for the Study of Teaching and Policy at the University of Washington (Grant # OERI-R308B70003); the National Science Foundation's Interagency Educational Research Initiative (Grant #s REC-9979863 & REC-0129421); the Atlantic Philanthropies, USA; and the William and Flora Hewlett Foundation.

Teacher Logs as a Tool for Studying Educational Process

For more than twenty years, research on teaching has had two central aims: to gather descriptive data on classroom teaching under different conditions of practice, and to estimate the effects of different teaching practices on student learning (Brophy and Good, 1986; Rowan, Correnti, and Miller, 2002). In studying these issues, educational researchers have generally used two strategies to gather data. The most common approach has been to send trained observers into classrooms to collect structured observational data, and more recently, to make video recordings of selected samples of lessons for later coding by experts. While these approaches are often viewed as the “gold standard” for classroom data collection, they are quite costly, and their use is typically confined to small scale studies (exceptions, however, are found in the use of video recording of teaching in the Third International Mathematics and Science Study [Hiebert, Stigler, Jacobs, et al., 2005] and the use of classroom observations in some large-scale, program evaluations commissioned by the U.S. Department of Education’s Institute for Education Sciences).

A second approach to collecting instructional data has been used by the National Center for Education Statistics (NCES) since the 1980’s. NCES large-sample surveys often include a small number of items on teacher surveys to measure teaching practices in American schools (e.g., the Schools and Staffing Survey, the National Assessment of Educational Progress, the Second and Third International Mathematics and Science Studies, the National Educational Longitudinal Studies of 1988 and 2002, the Early Childhood Longitudinal Study). Obviously, data from one-time surveys are less expensive to collect than observational or video-tape data, and are well-suited for large sample re-

search. However, researchers have questioned the accuracy and validity of these data (Mullens, 1995; Mayer, 1999).

This paper discusses a third (less frequently used) approach to gathering data on classroom instruction—the use of teacher logs. In this paper, we argue that teacher logs can be used in large-scale research on teaching and demonstrate data from teacher logs can be used to investigate methodological and substantive questions about classroom instruction. Our discussion focuses on data collected as part of the *Study of Instructional Improvement* (SII), conducted by the Institute for Social Research at the University of Michigan under contract with the Consortium for Policy Research in Education. SII collected data on approximately 150,000 reading/language arts and mathematics lessons carried out by approximately 2000 teachers working in 115 high-poverty elementary schools during the 2000-2001 to 2003-2004 school years.¹

The Problem

Classroom instruction is notoriously complex and difficult to measure (Jackson, 1990). Over the course of a nine-month academic year, the typical elementary school teacher will conduct 140 or more lessons in a given academic subject for the 20-30 students in her classroom, sometimes differentiating instructional activities by student or subgroups of students. Moreover, during any given lesson, a teacher's instruction will typically unfold along many different dimensions. For example, a teacher will normally cover several content objectives at different levels of cognitive demand during a single lesson in a subject like reading, working in several different behavior settings, using a variety of subject-specific instructional techniques. Although some features of classroom instruction

¹ For a description of the aims and methods of this study, see www.sii.soe.umich.edu.

are implemented repeatedly across the school year, many others are not, making instructional practice not only multi-dimensional, but also highly variable across days of the school year (Rogosa, Floden, and Willet, 1984).

Such complexity and variability present two problems for researchers. One problem occurs when survey researchers ask teachers to report on their teaching activities over an entire academic year in a survey administered near the end of that year—the most common data collection strategy in large-scale survey research. Here, teacher memory is the problem, in particular, the strong potential for inaccuracies in retrospective reports of the frequencies or rates at which particular teaching activities were undertaken (Burstein, 1995; Smithson and Porter, 1994). Variability in teaching creates a different problem, mainly for researchers conducting observations or making video recordings of instruction. Here, the problem is generalizability, that is, obtaining a sample of teaching observations that can be generalized to the universe of teaching events that unfolded over a nine month academic year. To the extent that teaching varies systematically across days of the school year, and especially if it involves many rare events, attempts to adequately sample teaching activities and reliably discriminate among different teachers' yearly patterns of instruction will require more in-class observations than all but the most well-funded studies can afford to collect.

These issues motivate our discussion of teacher logs. Logs (as typically administered) ask teachers to report on their instruction at the end of the school day, thus radically reducing the time period over which teachers must exercise recall. This, in turn, should increase teachers' reporting accuracy. Moreover, logs are really survey instruments, so they can be administered at much lower unit cost than classroom observations

or video-taping sessions, allowing data to be gathered on much larger samples of lessons, thereby improving the ability to generalize from a set of observations to the universe of teaching activities conducted over an entire academic year. Despite these advantages, logs are not without problems. As survey instruments, they are subject to errors in measurement due to social desirability, the response categories presented, and so on. Moreover, while frequent administration of logs can decrease coverage error and increase generalizability, these benefits come at the cost of increased respondent burden, which can increase survey non-response or lead to response bias if respondents develop time-saving (but inaccurate) patterns of filling out log surveys.

The purpose of this paper is to discuss some of the lessons learned about these issues as a result of administering teacher logs to a large sample of teachers during the *Study of Instructional Improvement*. We begin by discussing how the researchers conducting this study constructed the log, as well as the training procedures and incentives developed to increase response accuracy, minimize respondent burden, and improve response rates. We then discuss possible “survey errors” due to respondent and instrument error inherent in log-based measures of teaching (Groves, 1989). Here, we discuss how teachers’ log responses compared to the reports of trained observers and to the responses teachers made about their instructional activities on an annual survey completed at the end of the year. Next, we discuss different measurement models that can be used with log data and show how multi-level statistical models can be used to better understand psychometric properties of log-based measures. We conclude by discussing some of the main findings of our work about the nature of elementary school instruction and its consequences for student learning. Throughout the paper, we focus solely on the read-

ing/language arts log developed as part of the *Study of Instructional Improvement*. Readers interested in the mathematics log developed for this study can consult Rowan, Harrison, and Hayes (2004).

Logs and Log Administration

The reading/language arts logs administered as part of SII was a paper-and-pencil survey instrument containing over 100 items, mostly in checklist format, that teachers in 1st through 5th grades were asked to fill out at the end of a school day. Figure 1 is a copy of the log used in the study.

Figure 1 Here

In completing a log, teachers were instructed to report on the instruction provided to a single student in their reading class. To insure an accurate record of teachers' overall patterns of teaching, teachers' rotated log reports across a representative sample of eight students in their classes during three extended logging periods spaced evenly across the academic year. In this design, teachers who participated in all logging sessions with a complete roster of students would complete roughly 90 logs, or about 11 logs per student. However, because the elementary schools under study used many different instructional grouping arrangements for reading instruction, and because many students switched reading teachers mid-year, the average teacher in the sample completed about 39 logs over the course of the year.

The main purpose of the log was to gather data on several dimensions of instruction. The opening (or "gateway") section asked teachers to report on the amount of time spent by a focal student on reading/language arts instruction on the reporting day, as well

as the amount of emphasis given to each of the following topics: word analysis, concepts of print, oral or reading comprehension, vocabulary, writing, grammar, spelling, and research strategies. Then, if teachers checked that word analysis, comprehension, or writing was an emphasis for a student on a given day, teachers completed additional items (in the so-called “back end” of the log) about the specific content objectives that were taught to the student in that domain, the methods used to teach that content, and the tasks and materials the focal student engaged with that day.

Building on previous experience administering teacher logs for the *Panel Study on Income Dynamics* (Roth, Brooks-Gunn, Linver, and Hofferth, 2003), field staff conducting the *Study of Instructional Improvement* designed field administration procedures intended to both improve the accuracy of teacher responses to log items and assure adequate response rates. To improve response accuracy, field staff from the Institute for Social Research conducted a 45-minute training session for teachers before the first logging period of the year. This session introduced teachers to the definition of terms used in the log and taught teachers how to complete the log questionnaire. That session was followed by a suggested two-hour home study period, during which teachers were asked to study a glossary defining and illustrating the terms used in the log, and then by a one-hour, in-school, follow-up training session prior to the first logging period. Once logging began, teachers could call a toll-free phone number or ask local field staff to address any difficulties they were having with logging.

Also, an incentives plan was developed to increase teacher response rates. Using data on teachers' salaries nationwide, researchers calculated the average daily wage of teachers, and then offered payments to teachers based on the expectation that a single log

would take about 5 minutes to complete. Rather than use piece rate incentives, teachers were paid at the end of each logging period in which they logged. The actual incentive was \$150 per 6 week logging period if a teacher was logging for the full 8 students called for by the original design. In addition, field staff provided logging teachers with small gifts (coffee mugs, paper weights, pencils or pens) on a variable interval reinforcement schedule to further motivate log completion.

Overall, response rates for the reading/language arts log were quite high. About 90% of the teachers asked to log did so, and they completed 90% of the logs they were administered. Although teachers often struggled when they began logging, after about a week, they typically completed the reading language arts log in about five minutes time. Moreover, as the response rates demonstrate (and as teacher comments suggest), logging was not perceived as overly burdensome. Moreover, using logs was cost effective. A “back-of-the-envelope” calculation based on initial budgets and log administration data suggests a research cost of about \$27.50 per log administered, far less than the cost of conducting a single classroom observation or video-taping session, although more expensive than administering an one-time, annual survey to teachers.

Sources of Survey Error in Logs

These results demonstrate that it is possible to administer instructional logs to large samples of teachers over a prolonged period of time, and to achieve high response rates in doing so. However, an important question is whether or not the data gathered from logs is accurate. To examine this problem, we undertook several analyses to compare teachers’ log reports to: (a) log data collected by trained observers; and (b) teachers’ responses on an annual survey administered near the end of the academic year. From a “survey er-

ror” perspective (Groves, 1987), these analyses provide useful information about the different forms of measurement error that arise as different observers report on the same events and/or when different instruments are used to measure instruction as it unfolded over the course of a year.

Log vs. Observational Data

To examine observer bias in log use, we conducted a small study in eight public elementary schools where 31 teachers from various grades were pilot testing logging procedures. In the study, eight trained observers were sent two at a time into the classrooms of logging teachers on a given day during a three-month period, during which time, both teachers and observers completed logs. In this design, log reports for a single lesson were available from three individuals—two trained observers and a logging teacher. Using these data, Camburn and Barnes (2004: 54-60) conducted a number of statistical analyses to examine rates of agreement among teachers and observers, and observers themselves. They also used qualitative data from teacher interviews, narrative observation records, and observer reports to examine sources of disagreement among the teachers and observers who recorded data on the same lesson.

Camburn and Barnes’ (2004) statistical analysis focused on two types of teacher-observer agreement. The first, called a “gateway” match, occurred if and only if both teacher and observer completed the gateway section of the log so that both ended up in the same section of the back end of the log. Camburn and Barnes found that teachers and observers had appropriate gateway matches 81% of the time for instruction in word analysis, 90% of the time for reading comprehension instruction, and 87% of the time for writing instruction. A second type of match occurred at the “back end” of the log where

teachers completed checklists describing the specific content objectives, teaching practices, and instructional tasks and materials used in teaching a focal topic. Here, two kinds of matches were possible: matches where both teachers and observers checked an item during a lesson (a 1-1 match), and cases where both teachers and observers did *not* check the item during the lesson (a 0-0 match). Considering both matches simultaneously, Camburn and Barnes found the probability of teachers and observers producing a match on any item in the back end of the log to be about .73. However, since most items in the back end of the log were *not* checked during a given lesson, this high rate of matching resulted in part because 0-0 matches dominated the data set. Thus, Camburn and Barnes also examined the probability of 1-1 matches. Here, match rates were much lower — .22 for teacher-observer matches, and .41 for observer-observer matches. However, an important finding was that 1-1 match rates were much higher for items that were checked with highest mean frequencies in the data set (about .85 for the most frequently checked word analysis items, .77 for the most frequently checked comprehension items, and .75 for the most frequently checked writing items).

Using both match rate and qualitative data, Camburn and Barnes (2004) drew a number of conclusions about observer error in log surveys of instruction. First, the high match rates for gateway items and for frequently occurring items suggests that teacher logs are most accurate: (a) when describing instruction at a grosser (rather than finer) level of detail, and (b) for describing frequent (rather than rare) instructional practices. In addition, Camburn and Barnes concluded from their analysis of qualitative data that both teachers and trained observers made fallible reports of instructional practice. Indeed, a common error for both teachers and observers resulted from improper application of the

coding conventions enumerated in the log glossary. In comparison to trained observers, teachers' errors seemed to result when teachers overlooked quick, but routine, aspects of their teaching (e.g., correcting students' decoding errors as they read), while observer errors occurred when they were unable to see particular instructional acts, when particular instructional acts took place during very short lesson segments, or when observers misjudged teachers' intentions, particularly in terms of teachers' cognitive goals for students.

Logs vs. Annual Questionnaires

Camburn and Han (2006) conducted an additional analysis comparing teachers' log reports to their reports of instruction on the teacher questionnaire administered near the end of each school year. Here, data were drawn from the responses of 1,535 teachers in grades 1-5 who completed a reading/language arts log at least once during the *Study of Instructional Improvement* and who also completed an annual questionnaire the year they logged. By design, 24 items included on the language arts log were also included on the annual questionnaire, and by design, the wording of these questions was made as similar as possible on both instruments (although response formats differed).

Camburn and Han (2006) examined the amount and sources of *divergence* in teachers' responses to these 24 items across the log and questionnaire. The most important finding was that teachers uniformly reported higher frequencies of engaging in teaching practices on the annual questionnaires vs. the log. This is consistent with prior research on the correspondence between logs and questionnaires in educational settings (e.g., Burstein et al., 1995), but is inconsistent with research in other settings (e.g., Hoppe et al., 2000; Leigh, 2000; Leigh et al., 1998). Overall, the median difference in teachers' estimates of the frequency of teaching word analysis skills was about + 7.3

days/month (or nearly two days/week) on the questionnaire vs. the log, and about + 4 days per month (or one day a week) for comprehension and writing. The general tendency to over-report was more pronounced for female teachers, African American teachers, more experienced teachers, and teachers who individualized instruction. These latter findings constitute another form of observer bias (i.e., bias due to the characteristics of the respondent).

Psychometric Properties of Log Data

In this section, we discuss various measurement and statistical models that can be used to analyze log data once collected. Here, we discuss two challenges present in log data. First, log data consist mostly of dichotomously scored items, requiring analysts to move from statistical and measurement models based on the normal distribution to statistical or measurement models appropriate for categorical data. Second, log data (as collected in the *Study of Instructional Improvement*) are clustered, that is, hierarchically nested. In particular, daily observations from a single log are nested within students, students are nested within teachers, and teachers are nested within schools. As a result, analysis of log data requires analytic methods appropriate to clustered data.

Measurement Models for Log Data

The basic unit of instructional data collected during the *Study of Instructional Improvement* was a single log filled out by a teacher on a given day. In what follows, we call this lowest unit of analysis a “lesson.” Now, as teachers fill out the reading/language arts log, they have the potential to check over 100 separate items describing many different dimensions of instruction. To some extent, researchers will take an analytic interest in single items checked during a lesson—for example, the number of minutes on any given day

that a teacher taught reading/ language arts, or whether or not the teacher focused on a particular topic, say word analysis, reading comprehension, or writing. If that is the case, no measurement model need be applied to the data.

However, suppose an analyst wants to combine more than one item from the log to form a multi-item scale for a single, unidimensional, latent trait. When the log items to be included in this scale are dichotomously scored (0,1), an obvious approach to building multi-item scales at the lesson level is item-response theory (IRT), a form of latent trait analysis especially suited to the analysis of dichotomous items (for an accessible discussion of IRT, see Embretson and Reise, 2000). As an example, Rowan, Camburn, and Correnti (2004) used an IRT model to create a measure of the “cognitive demand” (or skill difficulty) of reading comprehension lessons taught to 3rd grade students on any given day. The reading/language arts log contained 12 dichotomous items that were assumed to describe this latent dimension of reading instruction, where some items were theorized in advance to index reading skills taught at a low level of cognitive demand (e.g., activating prior knowledge, previewing and surveying text), others were assumed to index more cognitively demanding skills (e.g., summarizing details in the text, sequencing information or events in the text), and still others were assumed to index highly demanding skills (e.g., analyzing/evaluating text, examining literary techniques). With lessons as the unit of analysis, Rowan, Camburn and Correnti (2004) used a one-parameter IRT model to construct a scale measuring the cognitive demand of reading instruction for each day in the data set. As expected, item difficulties from the estimated measurement model were in the theorized direction, the point bi-serial correlations of items to the total scale ranged from .56 to .28, and the scale had an estimated person reliability of .63.

In many cases, however, a data analyst might not be certain about the underlying traits being measured by some arbitrary (and possibly large) number of log items and will therefore want to explore the dimensionality of the data and reduce the number of dimensions measured to fewer than the number of items initially present in the data set. The most common tool used for this purpose is factor analysis. However, in most statistical software packages, the factor analysis subroutine requires a set of continuous observed variables and will yield incorrect results if the data are dichotomous, as with log data. One solution to this problem is to calculate the tetrachoric correlations between all item pairs and then factor analyze the resulting matrix as one would a matrix of Pearson correlations (using, for example, SAS PROC FACTOR). An alternative is to use one of the available statistical packages specifically designed for binary factor analysis, such as TESTFAC (Du Toit, 2003). Both procedures provide factor scores for a given case on each latent trait identified by the model.

These factor analyses can be quite informative. As an example, we have been interested in how word analysis was taught in particular lessons. The “back end” of the log contains a set of 9 items indexing two possibly distinct approaches to word analysis, four that index the strategy of sound blending, and five that index sound segmenting. A binary factor analysis showed that these items did in fact load on different dimensions, and as a result, we now use separate scores to index how much a lesson focused on one or the other approach to word analysis. We have obtained similar results in the areas of reading comprehension and writing, for example, reducing 31 items from reading section of the log to nine different, and theoretically meaningful, dimensions of comprehension instruction, and likewise for writing instruction. A problem, however, is that most of the scales

resulting from these factor analyses contain only 3-4 items and have low reliabilities.

It is worth noting, however, that many analysts will not be interested in constructing lesson-level measures of the sort just discussed and will instead simply want to aggregate data across lessons to the teacher level of analysis. Interestingly, although this procedure loses a great deal of information about how teaching practices vary over time for teachers, we have found that aggregation does little harm in terms of measurement. For example, when item level responses are aggregated into percentages for teachers, we have found that most percentages are nearly normally distributed and that linear factor analytic scores derived from these aggregated variables correlate in the range of .85 - .95 with the measures built up from the lesson level before being aggregated across teachers.

Generalizability Issues in Log Data

Although it is interesting to build measures of teaching at the lesson level, log data (as collected in the *Study of Instructional Improvement*) are highly clustered, with lesson-level measures nested within students, who are nested within teachers, who are nested within schools. In the early days of research on teaching, this kind of clustering presented formidable data analysis problems. Today, however, a number of statistical packages allow researchers to analyze clustered data, including HLM (Raudenbush, Bryk, and Congdon, 2000) and SAS PROC MIXED. In this section, we discuss how clustered log data can be analyzed using HLM in order to determine how many logs are needed to reliably discriminate patterns of teaching across various objects of measurement (e.g., students, teachers, or schools).

To begin, consider a very simple case where logs have been used to record the number of minutes of reading instruction across multiple lessons in a sample of teachers.

A simple two-level HLM for this continuous measure of teaching can be developed, where the level one model is $Y_{ij} = \beta_{0j} + e_{ij}$ and the level 2 model is $\beta_{0j} = \gamma_{00} + u_{0j}$. At level one, Y_{ij} is the number of minutes in reading instruction for a lesson occurring on occasion i taught by teacher j . In the model, this outcome is seen as varying randomly around the mean number of minutes for the teacher (across all observed lessons), where e_{ij} is an error term assumed to be normally distributed with mean 0 and variance (σ^2). At level 2 of the model, the mean lesson time for teacher j (β_{0j}) is seen as varying around the grand mean in lesson time for the whole sample (γ_{00}), plus a random teacher effect (u_{0j}) assumed normally distributed with mean 0 and variance τ_{00} .

This simple model can be used to examine the question of how many logs must be administered in order to reliably discriminate among teachers in their patterns of instructional time allocations. It is well-known, for example, that a researcher's ability to discriminate reliably among objects of measurement when measures are taken repeatedly is a function of three main factors: (1) the internal consistency of the measuring instrument; (2) the variance in "true score" measurements over time and across objects of measurement; and (3) the number of occasions on which measures are taken. If a single measurement tool is used, thereby controlling for errors of measurement, a simple expression describes the reliability with which we can measure β_{0j} , teachers' average lesson length. The formula is simply: $\alpha = \tau / [\tau + (\sigma^2/n_j)]$, where α is the reliability coefficient, τ is the variance among teachers in time allocations, σ^2 is the variance within teachers in time allocations, and n_j is the average number of observations of teachers. In general, the formula shows that reliability always increases as the number of observations increases. But, when variance among teachers (τ) is large, and variance across occasions (σ^2) is

small, only a few occasions of measurement are needed to reliably discriminate among teachers. Alternatively, as variance among teachers becomes smaller and/or variance across occasions increases, more observations will be needed to reliably discriminate among teachers. An important point is that if researchers have data on teachers' time allocations at multiple time points, and if they conduct the simple variance decomposition just described, they can simply plug different values of n_j into the formula just above to see how many observations they might need in future studies to achieve some desired level of reliability.

Although we illustrated this point with a two-level HLM, similar analyses can be conducted with more complex models, say HLM's with three or more levels, or non-linear HLM's where outcomes are dichotomous, count, or ordinal variables (for examples, see Raudenbush and Bryk, 2002: Chapter 10). In fact, we have conducted many such analyses in the course of our work (for one example, see Rowan, Camburn, and Correnti, 2004), and on the basis of these analyses we have reached some general conclusions. In general, we have found that for any given academic year, data from our logs cannot reliably discriminate among patterns of instruction experienced by students within the same classroom. As it turns out, this is not so much a result of inherent flaws in log data but rather occurs because the percentage of variance in instructional variables lying among students within classrooms is always tiny, while the percentage of variance in instruction lying within students over time is always very large. In fact, our data suggest that teachers do not much differentiate reading instruction among students, leading to the lack of reliability in estimating differences among students in instruction received.

By contrast, we are able to reliably discriminate patterns of instruction among teachers within the same school, and patterns of instruction across schools. In three-level HLMs, for example, the variance components we get in analyses suggest that our ability to reliably discriminate among teaching practices at the teacher level increases rapidly as the number of logs administered per teacher goes from 1 to about 10, increases more slowly from about 10 to 20 administrations, and then increases very little thereafter. So, collecting about 20 logs per year from teachers seems sufficient if the measurement goal is to reliably discriminate among teachers. Note, however, that the actual reliability obtained with 20 observations will depend on a number of factors. For example, it is very difficult to reliably discriminate among teachers when teaching events are rare, when the practice of interest varies greatly within teachers over time but very little across teachers, and when the dependent measure is unreliably measured (making σ^2 larger). In these cases, maximum reliabilities at about 20 observations can be as low as .60 or .70.²

Finally, it is worth noting that a researcher's ability to reliably discriminate among schools in patterns of reading instruction depends not only on the number of logs administered, but also the number of teachers sampled within schools, where increasing the sample of teachers completing logs will markedly improve between-school reliabilities. For example, if we are attempting to estimate teacher-level means in teaching practices at a single grade level in an elementary school, we will typically have only 3-4 teachers, and our school-level parameter estimates will have much lower reliability (.30 - .40) than if we look at patterns of teaching across all grade levels, where the presence of

² In fact, we have found that more than 20 logs will be needed in the domain of mathematics instruction, largely because many teaching events in this domain are rare, and because there is a great deal of day-to-day variation in mathematics teaching. Moreover, even in data from the *Study of Instructional Improvement* data, we often obtain reliabilities for teacher means in mathematics instruction well below the .70 mark discussed above.

15-20 teachers will produce school-level reliabilities on the order of .70 -.80 in most cases. Increasing the number of logs per teacher also will increase the reliability of school-level parameter estimates, but within any sized sample of teachers, increasing the number of logs beyond about 20 has virtually no effect on the reliability of school-level parameter estimates.

Substantive Findings with Log Data

The nested data provided by logs have allowed researchers working on the *Study of Instructional Improvement* to investigate variations in teaching practice at many different levels of analysis. In this section, we discuss what we have learned about variation in teaching across days of the week and year, about the extent and sources of variation in teaching practice among teachers within the same school, and about school-to-school variation in our sample. The most extensive treatment of these issues using log data from the *Study of Instructional Improvement* is Correnti (2005).

We begin by illustrating how log data illuminate the daily and yearly rhythms of reading/language arts instruction in American elementary schools. In general, log data from SII show that the frequency of reading/language arts lessons increases at a decelerating rate over the course of the academic year, reflecting a somewhat slow start and then the well-known November – April grind. Moreover, the frequency of lessons varies predictably across the school week. For example, reading/language arts lessons have been found to be less likely on Fridays, and on days just before and after holidays. Overall, the logs suggest that when both a student and his or her teacher are present in school, the student has about an 85% chance of having a reading language arts lesson on a given day,

the remaining days being given over to test preparation, field trips, assemblies, and other activities.

The logs show that reading/language arts instruction also varies predictably across grades. In 1st grade, reading lessons lasted about 90 minutes, but the amount of time given to the subject declined as students progressed through the grades, so that by 5th grade, the average reading/language arts lesson lasted only about 65 minutes in most schools. The content covered in reading/language arts and the level of cognitive demand of lessons also varied across grades. 1st grade teachers typically devoted about 40% of their lessons to word analysis, but this dropped to about 20% in second grade, and then to below 10% in 3rd grade and beyond. Meanwhile, the percentage of lessons devoted to reading comprehension and writing stayed about the same across grade levels—about 50% of lessons for reading comprehension and about 45% of lessons for writing, with the two subjects often taught together on the same day. However, while the amount of time devoted to these subjects stayed the same, the cognitive demand of lessons increased across as students progress through the grades. At higher grades, students tended to read and write longer and more complex texts, work on more demanding reading tasks, and engage in more planning and editing of their writing.

Despite these general trends, one of the most extraordinary findings from the *Study of Instructional Improvement* was the large variation that exists in teaching practice—even among teachers working in the same school and teaching at the same grade. In 1st grade, for example, where there is near universal agreement among reading experts that a heavy focus should be placed on the teaching of word analysis skills, our analytic models suggest that it would be very common to find two 1st grade teachers in the same

school one of whom focused on word analysis skills about a day a week and another who focused on this topic four days a week. Similarly large variations in teaching practice would be found across all content areas, with teachers in the same school at the same grade often varying by as much as 3-4 days a week in the percentages of lessons devoted to teaching reading and writing. Even more strikingly, we have found that very little of this variation is due to the average achievement levels of students in a classroom, or to variations in ethnic or socioeconomic composition, although there is a slight tendency for teachers with higher percentages of students with behavior problems to be less academically-focused. Moreover, variables indexing teachers' professional preparation (e.g., professional degrees, number of courses in different subjects, years of experience, pedagogical knowledge) have only tiny effects on teaching practices. In many ways, this extreme variability in teaching signals that schools remain "loosely coupled" organizations where teachers have considerable autonomy and function largely as curriculum brokers (Meyer and Rowan, 1978; Porter, 1989).

There are, however, some hopeful findings in our studies of teaching. The *Study of Instructional Improvement* was designed as a quasi-experiment that included groups of schools participating in three very different instructional reform programs, as well as a set of comparison schools not participating in these programs. One of the most striking findings of the study to date has been the extraordinarily large effects two of these reform programs had on the instruction occurring in schools. One of these programs, known as America's Choice (AC), was designed to foster a "literature-based" teaching regime that focused on writing as a means of improving students' reading comprehension. Analyses conducted by Correnti (2005) showed that teachers in AC schools were far more likely to

engage in this form of instruction than teachers in comparison schools, the odds ratio being about 4 (for AC vs. comparison schools) for the likelihood that on any given day, a reading comprehension lesson would also cover writing, and odds ratios ranging from 1.4 to 2.9 for other relevant indicators of this form of instruction. Similarly, the Success for All (SFA) program was designed to foster “skill-based” reading instruction, that is, reading lessons focused largely on basic reading comprehension skills. Analyses conducted by Correnti (2005) show that teachers in SFA schools were far more likely to engage in this form of instruction than were teachers in comparison schools, the odds ratios ranging from 4.4 to 1.4 for SFA vs. comparison schools on items indexing this form of instruction. Finally, in analyses not yet published, we are finding that these different forms of instruction produce gains in students’ measured reading comprehension, with skill-based reading instruction working better at the early grades (Rowan, Raudenbush, Correnti, et al., 2005), and literature-based instruction working better at later grades (Correnti, Rowan, and Camburn, 2003).

Conclusion

Overall, we have found that logs can be a cost-effective, reliable, valid way to measure instruction and examine the causal effects of instruction on student learning. Although the use of logs is more expensive than gathering data from annual questionnaires, our discussion suggests that log data are far more trustworthy than annual questionnaire data. Moreover, our discussion suggests that for many types of items—especially items measuring course grained features of instruction that occur frequently—logs can provide data that is nearly equivalent to what would be obtained by sending trained observers into classrooms. The data presented here further suggest that for most study purposes, ad-

ministration of somewhere around 20 logs (evenly spaced over the academic year) should allow researchers to reliably discriminate instructional practices in the area of reading/language arts across teachers and schools. Thus, using logs to gather data on instruction is a far less expensive way to gather adequately sized samples of reading instruction in large scale studies as compared to the used of observations. Finally, our analyses suggest that log data have strong construct validity, as shown by the effects of intervention programs on teaching, and by the effects of different kinds of teaching regimes on student learning. As a result, we believe that logs are a viable method of data collection in large scale research on teaching and that the use of logs should be expanded in the future.

References

- Brophy, J.E. & Good, T. (1986). Teacher behavior and student achievement. In M.C. Wittrock (Ed.), Handbook of research on teaching (3rd ed., pp. 328-375). New York: MacMillan.
- Burstein, L., McDonnell, L., Van Winkle, J., Ormseth, T., Mirocha, J., & Guiton, G. (1995). Validating national curriculum indicators. Santa Monica, CA: RAND.
- Camburn, E. & Barnes, C.A. (2004). Assessing the validity of a language arts instruction log through triangulation. Elementary School Journal, 105(1). 49-74.
- Camburn, E. & Han, S.W. (2006). *Factors affecting the validity of teachers' reports of instructional practice on annual surveys*. Madison, WI: Wisconsin Center for Education Research. Working paper of the Consortium for Policy Research in Education.

- Correnti, R. J. (2005). Literacy instruction in CSR Schools: Consequences of design specification on teacher practice. Dissertation: University of Michigan.
- Correnti, R.J., Rowan B., & Camburn, E. (2003). *Variation in 3rd grade literacy instruction and its relationship to student achievement among schools participating in Comprehensive School Reforms*. Paper read at the Annual Meeting of the American Educational Research Association, Chicago, IL, April.
- Embretson, S.E. & Reise, S.P. (2000). Item response theory for psychologists. Mahwah, NJ: Erlbaum.
- Groves, R.M. (1989) Research on survey data quality. Public opinion quarterly, 51, part 2, S156-S172.
- Hiebert J., Stigler J.W., Jacobs J.K., Givvin K.B., Garnier H., Smith M., Hollingsworth H., Manaster A., Wearne D., & Gallimore R. (2005) Mathematics teaching in the United States today (and tomorrow): Results from the TIMSS 1999 Video Study. Educational evaluation and policy analysis, 27 (2), 111-132.
- Hoppe, M., Gillmore, M., Valadez, D., Civic, D., Hartway, J., & Morrison, D. (2000). The relative costs and benefits of telephone interviews versus self-administered diaries for daily data collection. Evaluation review, 24(1), 102-116.
- Jackson, P.W. (1990). Life in classrooms. New York: Teachers College Press.
- Leigh, B. (2000). Using daily reports to measure drinking and drinking patterns. Journal of substance abuse, 12, 51-65.

- Leigh, B., Gillmore, M., & Morrison, D. (1998). Comparison of diary and retrospective measures for recording alcohol consumption and sexual activity. Journal of clinical epidemiology, 51(2), 119-127.
- Mayer, D. (1999). Measuring instructional practice: Can policymakers trust survey data? Educational evaluation and policy analysis, 21(1), 29-45.
- Meyer, J.W. and B. Rowan. (1978). The structure of educational organizations. In M.W. Meyer and Associates, Organizations and environments. San Francisco: Jossey Bass.
- Mullens, J.E. and Gayler, K.. (1999). *Measuring classroom instructional processes: Using survey and case study fieldtest results to improve item construction*. U.S. Department of Education. National Center for Education Statistics. Working Paper No. 1999-08.
- Porter, A.C. (1989). A curriculum out of balance: The case of elementary school mathematics. Educational researcher, 18(5), 9-15.
- Roth, J.R, Brooks-Gunn, J. & Linver, M.R. (2003). What happens during the school day? Time diaries from a national sample of teachers. Teachers college record, 105(3), pp 317-343.
- Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods. Thousand Oaks, CA: Sage.
- Rogosa D., Floden R., & Willet, J.B. (1984). Assessing the stability of teacher behavior. Journal of educational psychology, 76(6), 1000-1027.

- Rowan, B., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum in large-scale surveys: Insights from the Study of Instructional Improvement. Elementary school journal, 105, 75-102.
- Rowan, B., Correnti, R. and Miller, R. (2002). What Large-Scale, Survey Research Tells Us About Teacher Effects on Student Achievement: Insights From the *Prospects* Study of Elementary Schools. Teachers College Record, 104(8), pp 1525-1567.
- Rowan, B., D.M. Harrison, and A. Hayes. (2004) Using instructional logs to study mathematics curriculum and teaching in the early grades. Elementary school journal, 105, 103-127.
- Rowan, B., Raudenbush, S.W., Correnti, R., Schilling, S.G. & Johnson, C. (2005). *Studying "balance" in balanced literacy instruction: How different mixes of word analysis and text comprehension instruction affect first grade students reading achievement*. Paper prepared for research seminar on learning from longitudinal data, National Center for Education Statistics, May.
- Smithson, J. & Porter, A. (1994). Measuring classroom practice: Lessons learned from efforts to describe the enacted curriculum – the reform up close study (CPRE Research Report Series. No. 31). Madison, WI: Consortium for Policy Research in Education.