Using Teacher Logs to Measure the Enacted Curriculum:

A Study of Literacy Teaching in 3rd Grade Classrooms

Brian Rowan

Eric Camburn

Richard Correnti

University of Michigan

Abstract

In this article we examine methodological and conceptual issues that emerge when researchers measure the enacted curriculum in schools. After outlining key theoretical considerations that guide measurement of this construct and alternative strategies for collecting and analyzing data on it, we illustrate 1 approach to gathering and analyzing data on the enacted curriculum. Using log data on the reading/language arts instruction of more than 150 third-grade teachers in 53 high-poverty elementary schools participating in the Study of Instructional Improvement, we estimated several hierarchical linear models and found that the curricular content of literacy instruction: (a) varied widely from day-to-day; (b) did not vary much among students in the same classroom; but (c) did vary greatly across classrooms, largely as the result of teachers' participation in 1 of the 3 instructional improvement interventions (Accelerated Schools, America's Choice, and Success for All) under study. The implications of these findings for future research on the enacted curriculum are discussed.

That students are more likely to learn what they are taught in school than what they are not taught is clearly demonstrated in large-scale surveys of educational achievement where the overlap between what is taught and what is tested is measured. This research has shown repeatedly that students who are taught more of the curricular content appearing on an achievement test outperform students who are taught less of that content, controlling for other factors (Cooley & Leinhardt, 1980; Porter, Kirst, Osthoff, Smithson, and Schneider, 1993; Stedman, 1997). Thus, students' opportunities to learn specific topics in the school curriculum are both a central feature of instruction and a critical determinant of student learning.

The importance of curricular content to student learning has led researchers to become increasingly interested in measuring the "enacted curriculum" in schools, that is, the amount of instructional time devoted to teaching various strands and/or topics in the school curriculum (Porter, 2002). Indeed, measures of the enacted curriculum have become a central feature of many types of research in education, ranging from observational studies of classrooms (e.g., Fisher, Filby, Marliave, Cahen, Dishaw, Moore, and Berliner 1978; Knapp, Adelman, Marder, McCollum, Needels, Padilla, Shields, Turnbull, and Zucker, 1995) to large-scale surveys of American schooling (for reviews, see Brewer & Stasz, 1996; Rowan, Correnti, & Miller, 2002) to the high-profile international surveys of educational achievement (e.g., Schmidt, McKnight & Raizen, 1997; Westbury, 1992; Stedman, 1997).

Despite this trend, the procedures used to measure the enacted curriculum remain much as they were 2 decades ago. Large-scale surveys continue to administer teacher questionnaires once annually—a fallible procedure. Meanwhile, qualitative studies of instruction often conduct only a few observations, raising questions about how completely curriculum coverage was sampled over the school year. Across studies, widely varying lists of topics have been used to characterize the curriculum in U.S. schools, reflecting an arbitrarily chosen subset of curriculum objectives rather than the broader continuum of objectives to which students might be exposed in a given year. Other conceptual and methodological problems also plague the literature. There has been limited discussion of the advantages and disadvantages of using alternative strategies to gather data on the enacted curriculum and little sustained attention to how data on curriculum, once gathered, can be analyzed to produce tighter alignment among measurement procedures, data analytic strategies, and theoretical ideas.

What is needed to advance measurement of the enacted curriculum, we argue, is more attention to the theoretical foundations of research in this area, more discussion of the methodological challenges posed by theories, a fuller and more probing debate about alternative strategies for gathering data, and better empirical analyses showing how all of these issues can be addressed, if not resolved, in future work. To move forward on these goals, we discuss how a specific approach to measuring the enacted curriculum—instructional logs (or time diaries)—can address the challenges researchers face. We begin by outlining some key theoretical considerations that researchers confront when attempting to measure the enacted curriculum and list some conceptual and methodological problems that flow from these considerations. We then show how the use of instructional logs can address these problems and discuss a strategy for data analysis (involving the use of hierarchical linear models) that not only provides important information about the psychometric properties of log-based measures of the enacted curriculum but also allows researchers to test substantive hypotheses about this curriculum. Throughout this discussion, we focus on a study of literacy instruction in third-grade classrooms that was conducted in 53 high-poverty schools participating in the Study of Instructional Improvement.[i]

Background

Although focus of this article is on reading/language arts instruction in third-grade classrooms, our intention is not to contribute to research on literacy instruction *per se*. Instead, we seek to add to a broader area of research that Shavelson, Webb, and Burstein (1986) called "the measurement of teaching." This area has its origins in early observational studies of teaching (Medley & Mitzell, 1963) and was developed further during the heyday of process-product research on teaching (Rosenshine & Furst, 1973; Shavelson et al., 1986). In its earliest stages, research on the measurement of teaching was concerned mostly with how to conduct observational studies. By the 1980s, however, interest in the measurement of teaching spread to new research contexts. One development was the use of instructional logs (or time diaries) to gather data on classroom instruction, as was done in such studies as the Beginning Teacher Evaluation Study (Fisher et al., 1978) and research conducted at the Institute for Research on Teaching at Michigan State University (Floden, Porter, & Schmidt, 1980). Still later, an interest in the measurement of teaching spread to large-scale survey research when policy makers called for more and better survey data on instruction and annual surveys emerged as a primary means for gathering data on the enacted curriculum (Brewer & Stasz, 1996; Burstein, McDonnell, Van Winkle, Ormseth, Mirocha, and Guiton, 1995; Mayer, 1999; Mullens & Kasprzyk, 1999). Finally, interest in the measurement of teaching reached into the domain of teacher assessment, where new assessment strategies like portfolios, constructed-response items, assessment center exercises, and so on were developed (for a review, see Porter, Youngs, & Odden, 2001).

In a review of the literature on measurement issues in research on teaching written nearly 2 decades ago, Shavelson et al. (1986, p. 86) argued that "the domain of measurement in research on teaching is enormous[. It is] far too broad [to be treated] in any depth…[and] each topic within the domain is complex…" We agree, and as a result, limit this article to a discussion of how to measure a single construct in research on teaching (the enacted curriculum), within a single domain of the school curriculum (reading/language arts), at one grade level (third grade), using a single approach to measurement (teacher logs). This makes the scope of our discussion manageable and allows us to focus on a domain of measurement that has received sustained attention in prior research on teaching.

Within this domain, we focus on three issues. We begin by discussing theoretical conceptions of the enacted curriculum, the measurement problems associated with them, and the rationale we developed for using logs (as opposed to third-party observations or annual surveys) to address these problems. We then discuss analytic procedures that can be used to build measures of the enacted curriculum from log data, showing how these can be used to assess the psychometric properties of measures and to test substantive hypotheses drawn from the literature. We conclude with an empirical illustration of how such analyses can proceed, drawing on data from the log reports of approximately 150 third-grade teachers who recorded data on over 5000 days of reading/language arts instruction occurring in 53 high-poverty elementary schools during the 2000-2001 academic year.

Theoretical Foundations of Research on the Enacted Curriculum

The idea of measuring the enacted curriculum emerged slowly in research on teaching. Early observational studies largely ignored this construct, investigating the effects of teaching behaviors on student achievement in particular curricular domains without controlling for the potentially confounding effects of variations in curriculum coverage among teachers. This approach led to a great deal of criticism, however, and to a gradual recognition that measures of the enacted curriculum were central to research on teaching (for a review, see Shavelson et al., 1986, p. 54).

One result of this criticism was that early measures of the enacted curriculum were designed simply to control for the overlap between what was taught and what was tested in research on teaching (Cooley & Leinhardt, 1980; Husen, 1967). Here, the measurement strategy was to obtain a table of curriculum content for the achievement test being used in a study and then to ask teachers to check those content areas where instruction had been offered (during the period of study). Overlap measures almost always had statistically significant effects on achievement in this research, but the inclusion of such measures was never given much of a theoretical rationale in this research, serving instead as a kind of ad hoc control variable.

Soon, however, researchers linked measurement of the enacted curriculum to theoretical models of schooling. One linkage was to John Carroll's (1963) model of school learning and related ideas about time-on-task (see, e.g., measures of curriculum coverage used in the Beginning Teacher Evaluation Study [Fisher et al., 1978]). Now, the key task in measuring the enacted curriculum became quantifying students' exposure to curriculum content in terms of accumulated *time on a curriculum task* over some interval—usually the time elapsed between achievement tests. As this approach matured, measures of the time devoted to curricular content were based on a variety of response scales and built around a variety of aggregation procedures. But the main point remains. Even today, most research on the enacted curriculum conceptualizes schooling as a series of repeated (e.g., daily) exposures to instruction and takes as the key measurement problem to sample across days of instruction in order to produce an estimate of the overall amount or rate of exposure to particular elements of a curriculum that occurs during some fixed interval of time (usually an academic year).

As this research matured, researchers incorporated additional ideas about the school curriculum into their research. One development was the formulation of hierarchical conceptions of the curriculum. Here, the enacted curriculum was seen as having at least two dimensions worthy of measurement. The first was simply a list of the topics or objectives that constitute a given subject in the curriculum, with the lists generated varying greatly in terms of specificity and detail across studies. A second dimension was the cognitive complexity at which a given curriculum topic was taught (for a discussion of early developments in this line of work, see Shavelson et al., 1986, p. 54). Researchers attempting to measure this dimension usually imposed a hierarchical or developmental ordering on their measures, as illustrated, for example, in the well-known taxonomy of educational objectives developed by Bloom et al. (1956), or more recently, by Porter et al. (1993), who developed a set of coding categories for distinguishing levels of cognitive demand at which various topics within a subject are taught.

Sociologists of education developed a natural companion to these ideas—the idea of the curriculum as a differentiated opportunity structure. Here, researchers began to formulate questions about how access to curriculum unfolds over time for different groups of pupils (Barr & Dreeben, 1983; Oakes, Gamoran, & Page, 1992).

Clearly, if access to curriculum is an important determinant of achievement, then studying how teachers distribute opportunities to study different curriculum topics at varying levels of cognitive demand holds great promise for explaining how inequalities in learning emerge among students who enter schools with different social and academic backgrounds. Equally important, sociologists added a dynamic element to the study of curriculum coverage. For example, in at least some research the problem was not so much to develop summary measures of curriculum coverage over a given academic year as to measure the rate at which new material is introduced (see, e.g., the studies of pacing in the early reading instruction conducted by Barr & Dreeben [1983]).

A final set of ideas about the enacted curriculum emerged in the latest round of international comparisons of student achievement—especially work with the Third International Mathematics and Science Study (TIMSS) data set conducted by William Schmidt and colleagues (e.g., Schmidt, et al., 1997). Here, curricula are assessed in terms of their overall coherence or fragmentation—that is, the extent to which coverage is focused on a few key concepts or spread widely over many different (potentially loosely connected) topics. In this work, the U.S. mathematics curriculum has been called "a mile wide and an inch deep" and is argued to move at a snail's pace year after year, repeating many topics and introducing new material slowly. By contrast, the mathematics curricula in other nations are thought to be more focused and coherent, covering only a relatively few related topics in a given year and lacking the redundancy characterizing the U.S. curriculum.

Using Logs to Measure the Enacted Curriculum

The theoretical ideas present three challenges in building measures of the enacted curriculum: (1) choosing a cost-effective strategy for gathering data on the enacted curriculum, especially since the process under study unfolds over relatively long periods; (2) developing an analytic strategy to assess how patterns of content coverage are distributed across different objects of measurement, for example, lessons, students, and classrooms; and (3) developing measures of the enacted curriculum that are based on theoretical conceptions rather than ad hoc measures of the overlap between what is taught and what is tested.

Choosing a Data Collection Strategy

Three strategies have been used to collect data on the enacted curriculum: (a) third-party observations of classrooms; (b) log reports in the form of standardized checklists and/or other questionnaire items that are filled out frequently by teachers; and (c) questionnaires similar to logs in format but completed by teachers only once, usually near the end of an academic year.

By far, the cheapest data collection strategy is to administer annual questionnaires to teachers asking them to recall the topics they emphasized over an entire academic year. But many researchers have questioned the accuracy of teachers' responses to such surveys. For example, two widely cited studies were conducted to assess the consistency between content coverage measures based on annual teacher surveys and those derived from frequently administered teacher logs (Burstein et al., 1995; Porter et al., 1993). Burstein et al. (1995) found generally low correlations between these two methods for measuring content coverage, while Porter et al. (1993) reported correlations ranging from .80 to -.05 depending on the curriculum topic.

These findings are consistent with research in other social science fields comparing log and questionnaire responses. In fact, this broader body of research suggests some clear advantages of logs (or time diaries) over questionnaires administered only once. Apparently, single administration questionnaires that ask respondents to make retrospective self-reports of activities that transpired over relatively long periods of time suffer from two main problems. First, when the period over which reporting occurs is long, respondents can easily forget the behaviors in which they engaged. Second, when memory is fuzzy, respondents resort to estimating the frequencies of their behaviors. However, respondents use different estimation strategies, and as a result, two respondents with the same pattern of behavior often make very different retrospective reports. In fact, estimation has been found to be especially inaccurate in retrospective self-reports of two kinds of behaviors—those that rarely occur and those that occur frequently (see, e.g., Hilton, 1989; Hoppe et al., 2000; Leigh, 2000; Lemmens, Tan & Knibbe,1992; Lemmens, Knibbe & Tan, 1988; Sudman & Bradburn, 1982).

These findings are directly relevant to teachers' reports about curriculum coverage on annual surveys, where teachers are asked to estimate their curriculum coverage for an entire school year. On these surveys, teachers' undoubtedly face problems of recall that vary across curricular topics and lead to the use of different estimation strategies. Thus, it is not surprising to find low correlations among log reports and retrospective questionnaire reports in research on the enacted curriculum. To be sure, not all single-administration surveys will produce inaccurate data. Surveys asking for reports over short time periods (e.g., a day or week) should produce reasonably accurate reports. Moreover, various survey techniques can be used to increase respondents' recall and estimation accuracy (e.g., Menon, 1991). But annual surveys of curriculum coverage are known to be inaccurate, for all of the reasons just mentioned.

The limitations of annual surveys led us to consider two alternatives for collecting data on content coverage—third-party observations and teacher logs. Third-party observations are typically considered the "gold standard" for classroom research, so an interesting question is how teacher self-reports on logs compare to the reports of trained observers. There is only a modest amount of research on this topic, but existing studies have found acceptable convergence between observers' and teachers' reports of curriculum coverage for a given lesson (Burstein et al., 1995; Knapp et al., 1995; Porter et al., 1993).

An example of this kind of research is a small study conducted by Camburn and Barnes (2003) using pre-test data from the Study of Instructional Improvement. In this study, the reading/language arts lessons of 31 teachers were observed once by two trained observers, with both the observers and the teachers using the same logs to report on the content coverage and teaching activities that occurred during the observed lesson (Camburn & Barnes, 2003). In a sophisticated and wide ranging analysis of the correspondence between observers' and teachers' log reports, it was found that when match rates for teachers' and observers' reports were calculated using procedures typical of studies in this field, teachers' and observers' log ratings matched on 72%-84% of occasions. As in previous research, however, match rates varied across items. In fact, a striking finding of this study was that match rates were higher for curriculum topics and teaching practices that occurred more frequently in the data, and were lower for topics or practices that occurred less frequently.

Problems of Reliability in Observational Data

Camburn and Barnes (2003) also showed that both teachers' and observers' log reports were subject to measurement errors, although for different reasons. So, given the reasonably high teacher-observer match rates, it appears that teachers' log reports and third-party observations are both imperfect, but viable means for collecting data on the enacted curriculum. However, an additional consideration in choosing a data collection method in research is cost, and when data on curriculum coverage are being gathered over long periods (e.g., an entire academic year), it is important to consider how often data must be collected in order to produce reliable estimates of the phenomena under study. To the extent that many data-collection points are required, logs gain an advantage over third-party observations, because log data are less expensive to gather than equivalent data from third-party observers.

Consider how this works in practice. Suppose our goal is to gather data on curriculum coverage in classrooms over the course of an academic year in order to discriminate patterns of content coverage among teachers. The question is how many times we need to observe or administer logs to achieve this goal. It is well-known that the ability to discriminate reliably among objects of measurement when measures are taken repeatedly depends on 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, thus controlling for measurement reliability on occasions of measurement, a simple expression describes researchers' ability to reliably discriminate among objects of measurement (in this case, teachers). The formula is:

$$\alpha = \tau/[\tau + (\sigma^2/n_j)]$$

where α is a quantity between 0 and 1 expressing the ability to discriminate reliably among teachers in patterns of cumulative content coverage, τ is the amount of variance lying among teachers when this measure is averaged across occasions, σ^2 is the amount of variance lying within teachers across multiple occasions of measurement, and n_j is the average number of measurement occasions across all teachers.

The formula shows several things. First, a researcher's ability to discriminate across teachers increases as the number of occasions of measurement increases. But this reliability also depends on how much variation exists among teachers in overall patterns of content coverage (τ) , as well as how much variation there is from occasion to occasion in content coverage for each teacher (σ^2). If teachers vary greatly in cumulative content coverage, and if there is little occasion-to-occasion variance in coverage, relatively few observations are needed to discriminate among teachers. But if differences among teachers in cumulative content coverage are smaller and/or occasion variance is larger, relatively more observations are needed. In addition, the internal consistency of the measuring instrument affects the ability to discriminate reliably across teachers, largely by affecting occasion variance (σ^2). As the reliability of the measurement instrument decreases, σ^2 increases. Thus, holding all else constant, a lack of internal consistency in a measurement instrument requires researchers to increase the number of observations in order to discriminate reliably across objects of measurement (for an illustration, see Rowan, Raudenbush, and Kang, 1991; note also the correspondence here to the one-facet, nested, random, *g* study discussed in Shavelson & Webb, 1991).

To see how this works in practice, consider once again some pretest data collected for the Study of Instructional Improvement. Here, 40 elementary school teachers from varying grades completed an average of 40 reading/language arts logs over the spring semester. Using these data, we calculated the variance components just discussed: σ^2, the variance in a log item across occasions nested within teachers, and τ , the variance in that item across teachers. The item that varied most across teachers and least across occasions was the number of minutes teachers spent providing reading lessons. The proportion of variance among teachers for this item, also known as the intra class correlation ($\tau/[\tau + \sigma^2]$), was about .50. Figure 1 shows that, in order to obtain a measure of time spent on reading instruction that had a reliability of .90 for between-teacher discrimination, we would need to make about 14 observations of each teacher! Figure 1 also shows that when the proportion of variance among teachers is lower, as it was for all other items in the study, even more observations are needed to achieve high reliability. Clearly, these data suggest a clear advantage of logs over third-party observations in research on the enacted curriculum. When as many as 15-30 observations are needed to distinguish reliably among teachers, log reports provide a cost advantage over third-party observations.

[INSERT FIGURE 1 ABOUT HERE]

Objects of Measurement

To this point, we have been arguing that teacher logs provide reasonably accurate data about the enacted curriculum when they are filled out immediately after lessons and with enough frequency to reliably discriminate across objects of measurement. But this leaves an important question unanswered: What is the appropriate object of measurement in research on the enacted curriculum?

Research on teaching typically considers the teacher (or her/his classroom) to be the appropriate object of measurement for studies of the enacted curriculum. However, ideas about the curriculum as a differentiated opportunity structure call attention to the possibility that curriculum coverage differs across students in the same classroom. Indeed, a few studies of early reading achievement showed that when teachers used in-class grouping arrangements to teach early reading skills, patterns of curriculum coverage varied more widely among students within classrooms than among classrooms (Barr & Dreeben, 1983; Martin, Veldman, & Anderson, 1980). Sociological ideas about the enacted curriculum also suggest attention to the pacing of the curriculum across lessons, because differences in rates of curriculum coverage can be associated with growth in student achievement. However, the use of teachers (or classrooms) as unit of analysis has, in effect, precluded attention to students or lessons as objects of measurement in most research on the enacted curriculum.

In part, the focus on teachers as objects of measurement reflects the theoretical orientations of researchers in the field. In the early days of research on teaching, however, researchers also were hampered by methodological constraints in choosing a unit of analysis. In fact, the use of teachers as objects of measurement arose before the widespread availability of multilevel statistical computing packages and was required in earlier research because lesson- or student-level measures of content coverage could not be treated as statistically independent (for a discussion of this issue, see Shavelson et al., 1986; pp. 56-58). However, the widespread availability of computer programs for estimating multilevel statistical models has now eliminated this problem, allowing investigators to focus on multiple objects of measurement simultaneously.

In data analyses presented below, we estimate a series of three-level hierarchical regression models in order to study how the enacted curriculum unfolds at various (nested) levels of analysis—in our case, occasions, nested within students, nested within classrooms. In some of these analyses, our dependent variables will be dichotomous outcomes indexing whether or not a curriculum topic was taught on a given occasion; in other analyses the dependent variables will be scales measuring the cognitive demand at which curricular topics are taught. In analyzing the dichotomous data, we will use the hierarchical logistic regression models discussed by Raudenbush and Bryk (2002, Chap.10) in which the level-1 sampling model is a Bernoulli distribution and the hierarchical logistic regression model is:

[1] $$\eta_{tij} = \log [\varphi_{tij} / 1- \varphi_{tij}] = \pi_{0ij} + e_{tij},$$

[2] $$\pi_{0ij} = \beta_{00j} + r_{0ij}, \text{ and}$$

[3] $$\beta_{00j} = \gamma_{000} + u_{00j}$$

where φ_{tij} is the probability that a curricular topic of interest will be taught on occasion t, to student i, in classroom j; η_{tij} is the log odds that the topic will be taught on occasion t, to student i, in classroom j; π_{0ij} is student jk's log odds of being taught the topic; β_{00j} is the log odds that the focal topic will be taught in classroom j; and γ_{000} is the grand mean for the sample. In this model, σ^2 can be defined as Var (e_{tij}) or the unique variance in focal topic coverage across lessons for each student in the sample; τ_π as Var (r_{0ij}) or the variance in focal topic coverage across students within classrooms (assumed to have a mean of zero and to be normally distributed); and τ_β as Var(u_{00j}) or variance across classrooms in focal topic coverage (again assumed to have a mean of zero and be normally distributed).

As Raudenbush and Bryk (2002, Chap. 10) point out, we can use data on the variance components just discussed to estimate how reliably we can discriminate across students and teachers in patterns of content coverage. For example, our ability to reliably discriminate among students can be estimated as:

$$[4] \qquad \text{reliability of } \pi_{0ij\ (\text{estimated})} = \qquad \tau_\pi / [\tau_\pi + (\sigma^2/n_{jk})],$$

where n_{jk} is the number of lessons observed (on average) for students. And, our ability to discriminate reliably across teachers in focal topic coverage is:

$$[5] \qquad \text{reliability of } \beta_{00j(\text{estimated})} = \tau_\beta / (\tau_\beta + \{\Sigma[\tau_\beta + \sigma^2/n_{jk}]^{-1}\}^{-1}).$$

Thus, this basic model allows us to explore a number of measurement issues already discussed. First, using equations [4] and [5] we can assess our ability to discriminate reliably across objects of measurement in our study. Moreover, if we are willing to generalize from the variance component estimates in our data, we also can use equations [4] and [5] to estimate how the reliability in our study would change as the number of observations changes. All we need to do is use the existing values of σ^2, τ_π, and τ_β, and then insert different values of n_{jk} into equations [4] and [5]. This will give us the reliability coefficients we could expect for different numbers of observations, giving us a sense of how many observations we need to reliably discriminate patterns of content coverage at different levels of analysis.

Equally important, equations [1] through [3] can be expanded to include independent variables at all three levels of analysis. For example, at equation [1] of the model, we can include independent variables to assess the effect of the passage of time on the log odds that a curriculum topic will be taught during a given lesson, giving us some insight into the pace or unfolding of instruction. At equation [2], we can include variables to assess the effects of student characteristics such as prior achievement, gender, or socioeconomic status on the log odds of a topic being taught to a specific student, thereby examining whether (and on what basis) content coverage varies among students within classrooms. And, at equation [3] of the model, we can incorporate characteristics of teachers and their classrooms as independent variables in order to examine whether certain types of teachers, or those working with particular groups of students, display different patterns of content coverage.

In the analyses presented below, we also develop measures of curriculum coverage derived from scales that are assumed to be continuous, normally distributed variables. When these scales are the dependent variables, we are estimating a three-level hierarchical linear regression model as described by Raudenbush and Bryk (2002, pp. 228-230). This model differs from the one just presented only at equation [1]—the lesson level of analysis—where instead of estimating the log odds of a topic being taught on any given occasion, we are now using a scale score as the dependent variable (not a log odds), and σ^2 is assumed to be normally distributed with mean zero and equal variance within students. Otherwise, model assumptions and formulas for calculating reliabilities at higher levels of analysis are the same as above.

Thus, hierarchical linear (and logistic) regression models provide researchers with a flexible set of tools with which to study both substantive and psychometric problems arising in research on the enacted curriculum. They allow flexibility in the choice of objects of measurement, and they can be used to examine substantive hypotheses about why curriculum coverage varies across lessons, students, or classrooms. Moreover, the variance components in the data provide information needed to assess whether one can reliably discriminate across objects of measurement (above level 1 in the model), and, if one is willing to generalize from these variance components, to investigate how these reliabilities will change with different numbers of observations.

Theoretical Relevance of Measures

A final problem concerns how to represent substantive theoretical ideas about the nature of the school curriculum in the measurement process. Our interest in this issue stems from a dissatisfaction with measures focused on the overlap between the enacted curriculum and standardized achievement tests. The problem with overlap measures, as we see it, stems from the properties of achievement tests in U.S. society. Few would argue that standardized achievement tests present a desirable model for curricula, for they often lack such properties as curriculum coherence or focus—both of which have been seen as desirable properties of any curriculum in recent writing (e.g., Schmidt et al., 1997)—or they concentrate on only the narrowest slice of the curriculum, as at least some state tests do (see, e.g., Blank, Porter, & Smithson, 2001). In our view, an enacted curriculum that overlapped closely with such tests would not be a desirable curriculum. If it corresponded closely to the usual commercially produced achievement test, for example, it would probably be characterized by a low degree of both focus and coherence. And if it overlapped considerably with a narrowly focused state test, it would have the properties that critics of accountability programs complain about—an overly narrow focus on what is tested. Thus, overlap measures appear not to assess any theoretically derived or normatively desirable property of school curricula. Instead, they are simply measures of convenience that have been used to predict higher achievement on standardized tests.

We think it would be better to measure the enacted curriculum in terms of correspondence to normatively or theoretically derived ideas about the desirable properties of a curriculum. This raises the question, however, of what these desirable properties are. Our work takes two directions, both of which seek to assess molar properties of the curriculum as it is taught. One direction, not much discussed in this article, builds on ideas about the desirability of a focused curriculum emerging out of TIMSS work. Here, we use log data to assess the degree of focus in the reading/language arts curricula of schools, as measured by the number of topics taught during daily lessons. We also use log data to see if different teachers maintain distinctive curricular emphases. In these analyses, then, we assess not only the overall curricular focus (or lack of it) across different teachers, but also identify the distinctive foci of each teacher.

The second direction we are taking is discussed in this paper. Here, we examine the enacted curriculum from a developmental (or hierarchical) point of view. In the analyses that follow, for example, we characterize the curriculum along two dimensions. The first dimension characterizes the curriculum in terms of nine reading/language arts strands, including concepts of print; word analysis; reading comprehension; reading fluency; writing; vocabulary; grammar; spelling; and research strategies. Our view is that these are large areas (or domains) of the curriculum, that instruction in these strands is repeated across many grades, and that generally there is no linear sequence with which these strands are developed over time. We are therefore interested in analyzing strand coverage at the item level—asking, for example, what the likelihood is that an occasion, student, or classroom includes coverage of a curricular strand.

Within each strand, however, we argue that a second dimension of the curriculum can be identified. We call this dimension the developmental level of the curriculum, by which we mean the level of difficulty or cognitive demand of the skills being taught within a strand. To construct this measure, we capitalize on the fact that within particular curricular strands in our data, there is systematic variation in the frequency with which particular skills are taught. For example, in data on third grade reading, we have found that particular reading comprehension skills are taught with varying frequencies and that when these skills are arranged in descending order of frequency, there appears to be a natural progression in difficulty and cognitive complexity of the skills being taught. For example, the least cognitively demanding skills (e.g., those that help students activate prior knowledge or make predictions) are among the most frequently taught, while more demanding skills, such as those designed to help students actively comprehend text passages, to understand a larger story structure, or to compare and contrast multiple texts and literary styles are taught less frequently.

One way to test the assumption that the different rates at which these skills are taught can be used to produce a reliable scale measuring the difficulty of reading skills is to use the one-parameter item-response model originally developed by Rasch (1960) to model the occasion level data (for an accessible discussion of the family of item-response-theory [IRT] models, see Embretson & Reise, 2000). In using this approach, we recoded the original log items to denote whether or not one of the reading comprehension skills of interest was taught in a lesson. Then, using a Rasch model, we estimated the log odds that any item (i) would be covered on occasion (s) as a function of two parameters—the "trait" score (θ_s) of the occasion (a one-to-one function of the number of skills covered on a given day) and the item's level of difficulty (β_i), which is a direct function of the overall frequency of a given skill being taught in the sample of days, such that:

$$[6] \qquad \ln[P_{is}/(1-P_{is})] = \theta_s - \beta_i.$$

In this model, occasions that include more rarely taught skills receive a higher score (θ_s), a score that we argue indicates instruction that is more difficult in terms of the reading comprehension processes being taught. It should be noted that these data will have a good fit to a Rasch model only if occasions that cover the most difficult (i.e., least frequently taught) skills also cover easier (i.e., more frequently taught) skills. On the surface, this assumption might seem implausible, but we have found that the average lesson in our data covers about 5 reading comprehension skills, so the model is plausible.

We developed a similar measurement model for the written composition strand. Here, the skills included in the scale (from most to least frequently taught) were: writing practice; organizing ideas for writing; editing/capitals/punctuation; generating ideas for writing; sharing writing with others; editing word use/grammar/syntax; revising/refining/reorganizing writing; and revision through elaboration. Once again, the item frequencies seemed to correspond to a developmental pattern that begins with learning how to generate ideas, moves through revisions focused on grammar and syntax, and at the highest levels of difficulty involves progressive refinement and elaboration of written work.

In summary, our approach to measuring the enacted curriculum is designed to move beyond a focus on measuring the overlap between what is taught and what is tested in order to measure theoretically relevant dimensions of the curriculum. These dimensions include the degree and nature of curricular focus found in particular lessons and the level of difficulty of instructional content taught on particular days. Measuring these properties of the curriculum seems truer to the theoretical aims of measurement in research on the enacted curriculum than do measures of overlap.

Method

To demonstrate how these measurement approaches work in practice, we turn now to some analyses of data from the Study of Instructional Improvement. This is a study of the design, implementation, and instructional effectiveness of three of America's most widely-disseminated Comprehensive School Reform programs—the Accelerated Schools Program, America's Choice, and Success for All.

Schools

A major goal of this study is to examine instructional practices in schools working with these three programs. At the time of this writing, data relevant to this goal were available only for third grade classrooms in 53 schools that entered the study during its first year. These 53 schools were located in 33 districts in 11 states around the country. Fifteen were participating in the Accelerated Schools Program, 15 were in the America's Choice Program, 16 were in Success for All, and the remaining 7 were chosen as comparison sites because they were not participating in the three reform programs. At the time of data collection, 11 schools were implementing one of these programs for the first year, 21 schools were in their second year of implementation, 14 were in their third year, and the remaining 7 were the comparison sites. Overall, schools in the sample served a greater percentage of high-poverty students than would be expected in a representative sample of U.S. schools. For example, in the average school in the sample, about 73% of students were eligible for free and reduced-price lunches, and 76% were from minority backgrounds. In Fall of third-grade , the median reading level for students in the sample was the thirty-eighth percentile.

The Programs

At the time of data collection, two of the three programs under study (Success for All and America's Choice) had highly specified instructional designs in reading/language arts. Success for All was built around a 90-minute reading block composed of three timed segments—listening comprehension (20 minutes), reading instruction (55 minutes), and skills instruction (15 minutes), where the 55-minute reading block was designed to teach reading comprehension skills. The America's Choice program also had a distinctive instructional design, focusing (during early implementation) on the improvement of writing instruction through use of writer's workshops. By contrast, the Accelerated Schools Program lacked a well-specified instructional design for reading/language arts, working instead to help teachers internalize the imprecisely defined ideal of "powerful learning" in classrooms.

Log Data

All third-grade teachers in the 53 schools under study were asked to complete instructional logs, using the log shown in Appendix A. This log asks teachers to respond to simple checklists and other survey items to report on their instructional practices on a given day. In an initial section of the log, teachers report the time spent on reading/language arts instruction on that day and on the emphasis given to particular strands in the reading language arts curriculum. Then, if teachers checked one of the focal strands of the study (topics expected to be the most frequently taught and/or of special interest in the study), they were directed to complete additional items asking for more detail about curricular content and instructional activities.

Teachers are asked to complete logs during the Fall, Winter, and Spring periods of each academic year. On a given day, teachers reported on the instruction received by a single student in their class. A representative sample of eight target students from each third grade classroom was chosen at the start of the year. At the end of each logging day, teachers complete a log for one of the target students (randomly sampled from the eight), describing the reading instruction they received that day. In the data reported below, this procedure resulted in the collection of about 30-35 logs per teacher.

Measures of the Enacted Curriculum

Using log data, we developed two kinds of measures of the enacted curriculum. The first were measures of strand coverage, where the strands analyzed were reading comprehension and written composition (the two most frequently taught in the data). These measures were derived from question 4 in the log, where teachers reported the emphasis they placed on one of nine strands in the reading/language arts curriculum. If a teacher reported placing a major or minor focus on a topic on a log, we coded that strand as taught; if a teacher reported touching briefly on the strand or not teaching it, we coded the strand as not taught.

Within these strands, we used a Rasch scaling procedure to construct measures of the difficulty of reading and/or writing instruction occurring on a given day using items from section A of the log (for reading comprehension) and section B (for writing). These items are shown in Table 1. We also conducted an item analysis for these scales using the statistical package WINSTEPS v.3.07 (Linacre and Wright, 2000). The in-fit, out-fit, and reliability statistics from that analysis are also shown in Table 1. The reading difficulty scale had a reliability of .63; the writing difficulty scale had a lower reliability of .48.

Analytic Procedures

We arranged the data so that daily log reports on particular students (called lessons or occasions here) were nested within students, who were nested within teachers. In the analysis of strand data, this resulted in a sample of 5320 log reports on 668 third-grade students nested within 153 teachers. In the analysis of skill difficulty measures, we excluded days when the teacher or a student was absent, producing a sample of 4688 log reports nested within the same 668 third-grade students and 153 teachers.

The analysis proceeded in two steps. In the first step, we decomposed the variance in measures of the enacted curriculum into occasion, student, and teacher components and estimated the reliabilities of student- and teacher-level measures. In the next step, we included a set of independent (predictor) variables in the analysis. To shorten the discussion, the data sources and definitions of these independent variables are included in Appendix B. Our goal in including these variables was to examine substantive ideas about how instruction unfolds across the school year, to examine whether students' entry characteristics affected patterns of content coverage, and to examine relationships of teacher characteristics and classroom composition to patterns of content coverage. The school characteristic of interest in the analysis was the instructional improvement program in which a teacher participated.

All analyses were conducted using the computing package HLM 5.25 (Raudenbush et al., 2000). Because the strand data were measured as dichotomous variables the statistical model estimated was a three-level hierarchical logistic regression model, where the level 1 sampling model was a Bernoulli trial (see Raudenbush and Bryk, 2002, Chap. 10). When the dependent variables were the Rasch scale scores measuring the skill difficulty, the statistical model was a standard three-level hierarchical linear model (Raudenbush & Bryk, 2002, pp. 228-230).

Results for Strand Coverage

Table 2 shows the results of the variance-decomposition for data on reading comprehension and writing. The data show that the average classroom in the analytic sample had a 64% chance of a lesson covering reading comprehension and a 45% chance of a lesson covering writing. This model, based on 5,320 lessons, compares favorably with a similar fully unconditional model based on over 7,000 lessons, where the probability obtained for comprehension was .64 and .44 for writing. This is further evidence that the sub-sample of lessons used here is representative of the entire sample of lessons despite missing data on students and teachers. The dispersion statistics for both analyses show that the data conformed well to the Bernoulli sampling distribution. A surprising result in the analysis, however, was the lack of variance in strand coverage among students within the same classroom (although there was a great deal of variance among teachers). The reliability statistics reflected this, showing that we could not distinguish reliably among students' probabilities of receiving a reading comprehension or writing lesson (reliabilities were .001), although we could distinguish reliably among teachers' (reliabilities around .74).

[INSERT Table 2 ABOUT HERE]

Table 3 shows results after independent variables were entered into the model. At the occasion level, the log likelihood of lessons covering reading comprehension or writing did not vary across days of the week, but the likelihood that a lesson focused on comprehension or writing decreased over the school year. In addition, the log likelihood that reading or writing was taught on a given occasion was related to which other strands in the curriculum were taught on the same day. The significant relation of writing to comprehension (and vice versa), for example, suggests that these two strands were frequently taught together. Other strands were also positively or negatively related to the likelihood that comprehension or writing was taught. For example, lessons focused on word analysis, reading fluency, or vocabulary were more likely to cover reading comprehension but less likely to cover writing. A focus on concepts of print, grammar, and spelling were associated with an increased likelihood that writing was taught, but when grammar was taught, there was a decreased likelihood that comprehension was taught.

[INSERT TABLE 3 ABOUT HERE]

Given the absence of reliable variance among students in the data, it was not surprising to find only one statistically significant relation between a student predictor and measures of strand coverage—a small, positive effect of students' socioeconomic status on the likelihood that writing was taught. More surprising is that classroom composition variables such as average achievement, average SES, and percentage of White students were unrelated to strand coverage, the only exception being the positive effect of average SES on the likelihood that a lesson covered reading comprehension. Finally, teacher characteristics had some effects on strand coverage, with more experienced teachers being more likely to focus on both reading comprehension and writing, and professional development opportunities focused on teaching methods in reading/language arts during the past year being positively related to strand coverage in writing.

The most salient finding was the large and consistent differences among intervention programs on strand coverage. After entering indicator variables for each program into the regression models one at a time, we translated the estimates of intervention effects on log odds (shown in Table 3) into probabilities. If the average teacher in this sample were a Success for All teacher, she would have had an 85% chance of teaching a lesson focused on reading comprehension, whereas that chance was reduced to 53% for teachers in the America's Choice program, to 48% for teachers in the Accelerated Schools program, and to 41% for teachers in the control schools. The linear contrast among programs explained the majority of the 21% of variance in lesson coverage explained by the regression model.

Table 3 also showed large differences across programs in the probability that writing was taught. Using the procedure just discussed, we found that if the average teacher participated in the America's Choice program, she would have had a 65%

chance of focusing on writing. That probability declined to 39% for teachers in Success for All, 37% for teachers in control schools, and 36% for teachers in Accelerated Schools. Again, the linear contrast accounted for most of the 22% of explained variance in the analysis.

Skill Difficulty

Table 4 shows similar analyses of measures of skill difficulty in reading comprehension and writing. The bottom of Table 4 shows that approximately 62% of the total variance in skill level for reading comprehension was among occasions, less than 1% was among students, and about 38% was among teachers. The results for the writing scale showed even more occasion variance (80%), less than 3% of the variance among students, and 17% among teachers. One reason for the extraordinarily large occasion-level variance in writing, however, is the unreliability of the writing skills scale, which had a reliability of just .48 (suggesting that about half the occasion-level variance was due to measurement error).

[INSERT TABLE 4 ABOUT HERE]

Table 4 also showed that we could not reliably discriminate across students in skill levels taught in reading comprehension or writing (reliability = .02 for reading comprehension, and .18 for writing) but that we could discriminate reliably among teachers in their tendencies to teach at different skill levels in these strands (reliability = .90 for reading comprehension, and .75 for writing).

Reading comprehension. In the expanded regression models, the independent variables had different patterns of effects on skill difficulty for reading comprehension and writing. For reading comprehension, day of the week did not affect the skill level of lessons, but skill difficulty decreased over the school year. One reason for this was that lessons were less likely to focus on reading comprehension or writing as the year progressed (see Table 3). When there was no instruction on a strand on a given occasion, the difficulty of a lesson on that occasion was coded as the minimum score. Thus, as the probability that a strand was taught went down over time, the skill level at which that strand was taught also went down.

Table 4 also shows that the skill level of reading comprehension lessons was related to which strands (other than reading comprehension) were also taught on the same occasion. For example, when writing was taught, and when word analysis, concepts of print, reading fluency, and vocabulary were taught, reading comprehension lessons covered more difficult skills. By contrast, a focus on spelling decreased the skill level of reading comprehension lessons.

There was no evidence that teachers varied the skill level of reading comprehension lessons across students within their classrooms. However, skill levels in reading lessons varied systematically as a function of teacher (but not classroom composition) variables. Teachers with a master's degree in English and with more professional development in reading/language arts taught reading comprehension at a more advanced level.

Finally, Table 4 shows a large effect of instructional improvement programs on the average skill level of reading comprehension lessons, with teachers who participated in the Success for All Program showing a much higher average level than teachers in other programs. This effect approached .54 standard deviations (SDs) on the reading skills scale in comparison to America's Choice teachers, .50 SDs in comparison to teachers in the Accelerated Schools program, and .42 SDs in comparison to teachers in control schools.

Writing. The pattern of results was only slightly different for writing. Once again, the skill difficulty of writing lessons did not vary across days of the week, but declined as the year progressed. Again, this occurred because writing was less likely to occur (see Table 3). Moreover, Table 4 shows that the skill level of writing lessons was higher on occasions when teachers also focused on reading comprehension, grammar, spelling, research strategies, and concepts of print. Writing was taught at a lower skill level, however, when a teacher also focused on reading fluency, vocabulary, and/or word analysis.

Table 4 also shows that teachers did not vary the skill level of writing lessons among students in their classrooms or across classes with different ethnic, socioeconomic, or achievement composition. Moreover, in the case of writing instruction, only a single teacher characteristic—the amount of professional development a teacher received in reading/language arts—affected the skill level of writing lessons.

There was a significant effect of school improvement programs on writing instruction, with teachers in the America's Choice Program providing writing lessons at higher skill levels than teachers in other programs. Before entering a variable measuring a teachers' exposure to professional development in teaching methods into the regression models, the effect for America's Choice teachers versus teachers in Success For All and the Accelerated Schools Program was statistically significant, with effect sizes of .19 and .20 SDs, respectively, on the writing skills scale. Moreover, there were large difference between America's Choice teachers and comparison teachers in these analyses, but the small number of teachers in the comparison group produced a high standard error of measurement, reducing the statistical significance of this comparison. In the final conditional model shown in Table 4, however, America's Choice teachers were only significantly different from teachers in the Accelerated Schools Program, the effect size here being .17 SDs.

Discussion

Several themes emerged from these analyses. Some of these are related to psychometric issues in the measurement of the enacted curriculum, but these psychometric issues have important implications for theoretical ideas as well. Summarizing the data analyses, we see the following main points:

1. **The largest amount of variation in the enacted curriculum occurs at the occasion level, suggesting that teachers vary the content and difficulty of the skills they teach widely from day to day.** This has important psychometric impli-

cations for researchers using third-party observations or logs to measure the enacted curriculum, because more observations are typically required for reliable measurement when occasion variance is high. Substantively, our analyses show that it is possible to analyze occasion variance in content coverage and skill levels and thereby produce interesting findings about time trends in curriculum coverage and about relationships among curriculum topics.

2. **There is little evidence of differentiation in either the amount or skill level of reading comprehension or writing instruction received by students in the same classroom over the course of a year.** An important issue arising from this finding is that it would never be possible to reliably discriminate across students as objects of measurement under these conditions, no matter how many times students are observed receiving instruction! This is important because some researchers have aggregated occasion-level data on instruction to the student level, and used such data to detect reliable differences in curriculum coverage among students. Our data, however, raise the possibility that these reliable differences could have been a function of sampling error due to observing students on different occasions. For this reason, we advise researchers to take into account variance in instruction across occasions before using students as objects of measurement in research on the enacted curriculum.

3. **Even with large occasion variance in content and skill coverage, it is possible to reliably discriminate among teachers in patterns of curriculum en-**

actment. However, our data suggest as many as 20 to 30 observations per teacher might be needed to obtain reliable estimates. When combined with the finding of small student-level variance in content coverage, our data also suggest that it might be safe to measure teachers' content coverage without controlling for the types of students being taught. Moreover, if one is willing to ignore variation in content coverage arising from occasions, it might even be safe to produce teacher-level measures from log data by aggregating over all occasions and working with summary data, especially when logs are filled out on numerous occasions. In fact, we explored a variety of measurement strategies that aggregated data to the teacher level without controlling for variation across students and occasions and found that the correlations among scale scores assigned to teachers using HLM estimates corrected for student and occasion error and aggregate analyses not correcting for this error was always in the range of .85-.95.

4. The effects of intervention programs on the enacted curriculum were large and consistent with the intended designs of the interventions under study. In a study concerned with examining differences in the enacted curriculum across intervention programs, this is an important finding. But it is also important for educational policy, for it suggests that curriculum coverage is an alterable variable and that intervention strategies such as the ones used by the school reform programs we are studying can bring teachers' content coverage decisions more into line with planned curricula.

Summary and Conclusion

In summary, this paper shows how a richer theoretical conceptualization of the enacted curriculum can be developed and how data can be analyzed when teacher logs are used to collect data on teaching. In particular, we advocate moving beyond conceptions of the enacted curriculum as the overlap between what is taught and what is tested in order to measure more theoretically relevant properties of the curriculum and the use of hierarchical regression procedures to model variation in curriculum enactment occurring across occasions, students, and teachers.

The analyses presented here show the promise of this approach. Especially important, in our view, were the findings of program effects on teachers' curricular decision making. These findings suggest that teachers have less autonomy in enacting the curriculum than popular images of schools as loosely coupled systems and teachers as curriculum brokers suggest. In fact, our data suggest that intervention programs can have powerful effects on the enacted curriculum in American schools and that curriculum coverage in U.S. classrooms can, after all, be treated as an alterable variable in discussions of educational reform.

All of this reinforces our call for better theory, measurement, and analysis of the enacted curriculum. If curriculum is an alterable variable, then more and better research is needed on the properties of different curricula, on the ways in which curricula are enacted in classrooms, and on the effects of curricula on student learning. We invite other researchers to use the kinds of data collection and analytic strategies we developed to investigate these issues, and we stand open to suggestions for how our own work in this area might be improved.

References

Barr, R. & Dreeben, R. (1983). How schools work. Chicago: University of Chicago Press.

Blank, R., Porter, A.C., & Smithson, J. (2001). New tools for analyzing teaching, curriculum, and standards in mathematics and science. Washington, DC: Council of Chief State School Officers.

Bloom, B. & others. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York: Longmans Green.

Brewer, D.J. & Stasz, C. (1996). Enhancing opportunity to learn measures in NCES data. Santa Monica: RAND.

Burstein, L., McDonnell, L., Van Winkle, J., Ormseth, T., Mirocha, J., & Guiton, G. (1995*)*. Validating national curriculum indicators. Report prepared for the National Science Foundation. Santa Monica, CA: RAND.

Camburn, E. & Barnes, C. (2003). Assessing the validity of an instruction log through triangulation. Submitted to Elementary School Journal.

Carroll, John B. (1963). A model of school learning. Teachers College Record, 64, 723-733.

Cooley, W.W. & Leinhardt, G. (1980). The instructional dimensions study. Educational Evaluation and Policy Analysis, 2, 7-25.

Embretson, S.E. & Reise, S.P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Earlbaum Associates.

Fisher, C., Filby, N., Marliave, R., Cahen, L., Dishaw, M., Moore, J., & Berliner, D. (1978). Teaching behaviors, academic learning time, and student achievement. (Phase III-B, final report) Beginning teacher evaluation study. San Francisco: Far West Laboratory.

Floden, R.E., D.J. Freemen, A.C. Porter, & Schmidt, W.H.. (1980). Don't they all measure the same thing? Consequences of selecting standardized tests. In E.L. Baker & E. Quellmalz (Eds.), Design, analysis, and policy in testing and evaluation. Beverly Hills: Sage.

Hilton, M. (1989). A comparison of a prospective diary and two summary recall techniques for recording alcohol consumption. British Journal of Addiction, 84, 1085-1092.

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.

Husen, T. (Ed.) (1967). International study of achievement in mathematics: Comparison of twelve countries (Volumes 1 and 2). New York: John Wiley.

Knapp, M., Adelman, N., Marder, C., McCollum, H., Needels, M.C., Padilla, C., Shields, P.M., Turnbull, B.J., & Zucker, A.A. (1995). Teaching for meaning in high poverty 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.

Lemmens, P., Tan, E., & Knibbe, R. (1992). Measuring quantity and frequency of drinking in a general population survey: A comparison of five indices. Journal of Studies on Alcohol, 53, 476-486.

Lemmens, P., Knibbe, R., & Tan, F. (1988). Weekly recall and diary estimates of alcohol consumption in a general population survey. Journal of Studies on Alcohol, 49, 131-135.

Linacre, J.M. & Wright, B.D. (2000). A user's guide to WINSTEPS rasch model computer program. Chicago, IL: MESA Press.

Martin, J., D. Veldman, & Anderson, L.M. (1980). Within-class relationships between student achievement and teacher behaviors. American Educational Research Journal, 17(4), 479-490.

Mayer, D.P. (1999). Measuring instructional practice: Can policymakers trust survey data? Educational Evaluation and Policy Analysis, 21(1), 29-46.

Medley, D.M. & Mitzell, H.E. (1963). Measuring classroom behavior by systematic observation. In N.L. Gage (Ed.), Handbook of Research on Teaching. Chicago, IL: Rand McNally.

Menon, G. (1991). Judgements of Behavioral Frequencies. Unpublished doctoral dissertation. University of Illinois at Urbana-Champaign.

McLaws, M.L., Oldenburg, B., Ross, M., & Cooper, D. (1990). Sexual behavior in AIDS-related research: Reliability and validity of recall and diary measures. Journal of Sex Research, 27(2), 265-281.

Mullens, J. & Kasprzyk, D. (1999). Validating item responses on self-report teacher surveys. Washington, DC: U.S. Department of Education. NCES working paper.

Oakes, J., Gamoran, A., & Page, R. (1992). Curriculum differentiation: Opportunities, outcomes, and meanings. In P. W. Jackson (Ed.), Handbook of research on curriculum (pp. 570-608). New York: Macmillan.

Porter, A.C. (2002). Measuring the content of instruction: Uses in research and practice. Educational Researcher, 31(7), 3-14.

Porter, A.C., Kirst, M.W., Osthoff, E.J., Smithson, J.L., & Schneider, S.A. (1993). Reform up close: An analysis of high school mathematics and nce classrooms. Madison, WI: Wisconsin Center for Educational Research.

Porter, A.C., P. Youngs, & Odden, A. (2001). Advances in teacher assessments and their uses. In, V. Richardson (Ed.), Handbook of research on teaching (4th ed.). New York: Longman.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Raudenbush, S.W. & Bryk, A. (2002). Hierarchical linear models: Applications and data analysis methods. Thousand Oaks, CA: Sage.

Raudenbush, S.W., A. Bryk, Y.F. Cheong, & Congdon, R. (2000). HLM 5. Chicago: Scientific Software Internation.

Rosenshine, B. & Furst, N. (1973). The use of direct observation to study teaching. In R.M.W. Travers (Ed.), Second handbook of research on teaching. Chicago: Rand McNally.

Rowan, B., R. Correnti, & Miller, R.J. (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), 1525-1567.

Rowan, B., S.W. Raudenbush, & Kang, S.J. (1991). School climate in secondary schools: A multilevel analysis. In S.W. Raudenbush and D.J. Willms (eds.), Pupils, classrooms, and schools: Multilevel studies in education from an international perspective. New York: Academic Press.

Schmidt, W.H., C. McKnight, & Raizen, S. (1997*).* A splintered vision: An investigation of U.S. science and mathematics education. Dordrecht, The Netherlands: Kluwer.

Shavelson, R.J., N.M. Webb, and L. Burstein. (1986). Measurement of teaching. In, M.C. Wittrock (Ed.), Handbook of research on teaching (3rd edition). N.Y.: MacMillan.

Shavelson, R.J., and Webb, N.M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage Publications.

Stedman, L.C. (1997). International achievement differences: An assessment of a new perspective. Educational Researcher, 26(3), 4-14.

Sudman, S. & Bradburn, N.M. (1982). Asking questions. A practical guide to questionnaire design. San Francisco: Jossey Bass.

Westbury, I. (1992). Comparing American and Japanese achievement: Is the United States really a low achiever? Educational Researcher, 21 (5), 18-24.

Appendix A

Study of Instructional Improvement Language Arts Log – Page 1

See language arts log at the end of this paper

Appendix A (Continued)

Study of Instructional Improvement Language Arts Log – Page 2

See language arts log at the end of this paper

Appendix A (Continued)

Study of Instructional Improvement Language Arts Log – Page 3

See language arts log at the end of this paper

Appendix A (Continued)

Study of Instructional Improvement Language Arts Log – Page 4

See language arts log at the end of this paper

Appendix B

Description of Independent Variables Included in Prediction Models

Variable	Description
Teacher/Classroom:	
Intervention program	Set of 4 dummy variables indicating which intervention program, if any, the teacher's school is involved in. For example, America's Choice dummy is coded "1" if teacher is in school participating in America's Choice, "0" otherwise. Of the 53 schools, 15 participated in AC and ASP, 16 participated in SFA and 7 were Comparison schools.
Classroom:	
Average fall achievement	Average fall scale score for the eight target students in each classroom on the Terra Nova Reading Sub-component.
Average SES	Average socioeconomic status for eight target students in each classroom.
% of White students	Percentage of students in each classroom who are white.
Teacher has master's degree	Dummy variable indicating whether or not teacher has obtained their Master's degree.

	101 out of 153 teachers in the sample, or roughly two-thirds, had obtained their Master's.
Missing data on master's degree	Dummy variable indicating if the teacher's degree information was missing from the teacher questionnaire. 8 respondents were missing.
Self-contained	Dummy variable indicating if teacher taught in a self-contained classroom. 125 of the 153 teachers in the sample taught in a self-contained classroom.
Missing data on self-contained	Dummy variable indicating if teacher's role at the school was missing. Only 5 teachers had missing information about their role at the school.
Amount of professional development on teaching methods	A Rasch IRT scale score comprised of four items, including (1) the number of professional development sessions the teacher participated in which focused on teaching methods (1=none, 4=8 or more sessions) (2) the amount of time and effort the teacher devoted to improving their knowledge of the writing process (1=none, 7= a great deal) (3)

	the amount of time and effort devoted to extending their knowledge about different reading comprehension strategies such as KWL or reciprocal teaching (1=none, 7=a great deal) and (4) how often they worked with other faculty or staff developing thematic units or other approaches to integrating instruction across curricular areas (1=never, 5=more than 10 times). A high score on this measure indicates that the teacher's professional development focused highly on methods for teaching literacy. The scale has a Rasch reliability of .66.
Years experience	The number of years of experience the teacher has in any school.
Student:	
Student's fall achievement	Student's fall achievement scale score on the Terra Nova Reading Sub-component.
Male	Dummy variable coded "1" if student is male, "0" if student is female. 330 out of 668 students in the sample, or 49%, are male.
White	Dummy variable coded "1" if student is

	white, “0” otherwise. 192 students, or 28.7% of the 668 students are white.
SES	Continuous scale indicating the student’s socioeconomic status. This composite was determined by a set of five variables – mother’s professional status, mother’s level of education, father’s professional status, father’s level of education and household income.
Engagement (teacher rating)	Scale of 11 items where the teacher was asked to rate for each target student whether the student 1) is eager to learn 2) usually pays attention in class 3) completes school work in an organized way 4) works well independently 5) wants to do well in school 6) keeps personal belongings organized 7) works hard in school 8) persists when work is difficult 9) usually completes work on time 10) uses free time constructively 11) works carefully and methodically. Items were all on a 4 point scale from strongly disagree to strongly agree. The scale accounts for 70% of the joint variance in these items and has an alpha reliability of .96. Higher scores on the scale indicate

higher student engagement.

Occasion:

Log date

Count of the number of days since the first logging day. Values range from 1 to 235.

Day of the week

A set of five dummy variables coded “1” if day was Monday, “0” otherwise for Monday; coded “1” if day was Tuesday, “0” otherwise for Tuesday; etc.

Holiday

Dummy variable indicating if day was a holiday or occurred immediately prior to a holiday weekend. Nine out of 114 days sampled were coded as being affected by a holiday, and they include, Halloween, the day before Thanksgiving, Valentine’s day, Friday before President’s day weekend, Friday before St. Patrick’s day weekend, Friday before Easter weekend, Monday after Easter weekend, Friday before Memorial day weekend, and Memorial day.

Curriculum strand focus	Nine variables indicating the degree of focus (for each lesson) on each of the nine curriculum strands on the log, including: 1) comprehension 2) writing 3) word analysis 4) concepts of print 5) reading fluency 6) vocabulary 7) grammar 8) spelling and 9) research strategies. A Score of “0” indicates strand was “not a focus”; “1” indicates strand was “touched on briefly”; “2” indicates strand was a “minor focus”; and “3” indicates strand was a “major focus”.

Table 1. Reading Comprehension and Writing Item Statistics

		Rasch Model Statistics			
Items	% lessons (when topic strand focused on)	Item Difficulty	Infit	Outfit	Point-Biserial Correlation
Comprehension:					
Activating prior knowledge	69	-2.02	.98	1.00	.44
Previewing, predicting, surveying text	62	-1.54	.98	1.11	.47
Summarizing important details in text	55	-1.08	.99	1.00	.49
Self-monitoring for meaning	43	-.38	1.02	1.05	.48
Sequencing information/events in text	37	.02	.98	.94	.51
Using visualization/imagery to understand text	33	.34	.90	.85	.55
Using concept maps/frames	31	.50	1.03	1.03	.47
Identifying story structure	30	.51	.96	1.05	.50
Analyzing/evaluating text	29	.62	1.01	1.05	.47
Comparing/contrasting information in text	27	.79	1.01	1.07	.46
Using charts or visual aids	26	.86	1.02	.98	.46

Examining literary techniques	21	1.38	1.09	1.26	.37
Writing:					
Writing practice	63	-.71	1.21	1.34	.28
Organizing ideas for writing	62	-.63	1.01	.94	.46
Editing capitals, punctuation, or spelling	55	-.26	.80	.75	.62
Generating ideas for writing	54	-.19	1.23	1.29	.29
Sharing writing with others	48	.21	1.11	1.15	.38
Editing word use, grammar, or syntax	45	.39	.84	.77	.60
Revision of writing:					
Refining or reorganizing	44	.49	.92	.93	.53
Elaboration	41	.70	.88	.81	.56

Table 2. Three-Level HGLM Variance Decomposition for Reading Comprehension and Writing (n=5320 lessons)

	Comprehension	Writing
Estimated probability of focusing on strand[a]	.64	.45
Dispersion Statistic	.944	.948
Variance Component:		
Student	.001	.001
Teacher	.916	.817
Reliability:		
Student	.001	.001
Teacher	.743	.740

[a]The estimates shown here are based on the unit-specific model, where coefficients were converted to a probability using the following equation $1/(1+e^{-(coef.)})$

Table 3. HGLM Estimates of the Log Odds that a Lesson Will Focus on Reading Comprehension and Writing (n=5320 lessons)

	Comprehension		Writing	
Variables	Coefficient	SE	Coefficient	SE
Intercept	.65	.09	-.27	.08
Teacher:				
America's Choice	-1.20***	.24	.86***	.23
Accelerated School's	-1.22***	.21	-.10	.21
Comparison	-1.67***	.35	-.12	.35
Has Master's degree	.04	.20	-.29	.20
Missing data on Master's degree	.71	.38	.19	.37
Amount of professional development on teaching methods	-.04	.10	.30**	.09
Years experience	.02**	.009	-.02*	.008
Classroom:				
Average Fall Achievement	-.004	.004	.003	.004
Average SES	.77**	.24	-.10	.22
Percentage of white students	-.07	.30	.07	.28
Self contained	.02	.30	-.26	.27
Missing data on self contained	.13	.61	.06	.54
Student:				
Fall achievement	.001	.002	.00	.001
Male	-.10	.09	.14	.09

White	.17	.14	.00	.14
SES	-.01	.07	.15*	.07
Engagement (Teacher rating)	.04	.06	-.03	.06
Occasion:				
Log date	-.01***	.001	-.005***	.001
Monday	.12	.13	-.04	.13
Tuesday	.19	.13	.17	.13
Wednesday	.17	.13	.16	.13
Thursday	.12	.13	.22	.13
Holiday	-.36*	.14	.15	.16
Focus of Lesson:				
Comprehension	--	--	.45***	.03
Writing	.46***	.04	--	--
Word analysis	0.21***	.05	-.16**	.05
Concepts of print	.04	.08	.39***	.07
Reading fluency	.60***	.05	-.26***	.05
Vocabulary	.22***	.05	-.12*	.05
Grammar	-.16**	.06	.41***	.06
Spelling	-.06	.05	.51***	.05
Research strategies	-.05	.07	.12	.07
Summary statistics:				
Variance Component:				
Student	.002		.001	

Teacher	.728	.641
% Reduction in teacher variance from table 1	21%	22%

p<.05 ** p<.01 *** p<.001

Table 4. HLM Estimates of Lesson, Student, and Classroom Effects on Skill Difficulty in Comprehension and Writing (n=4688)

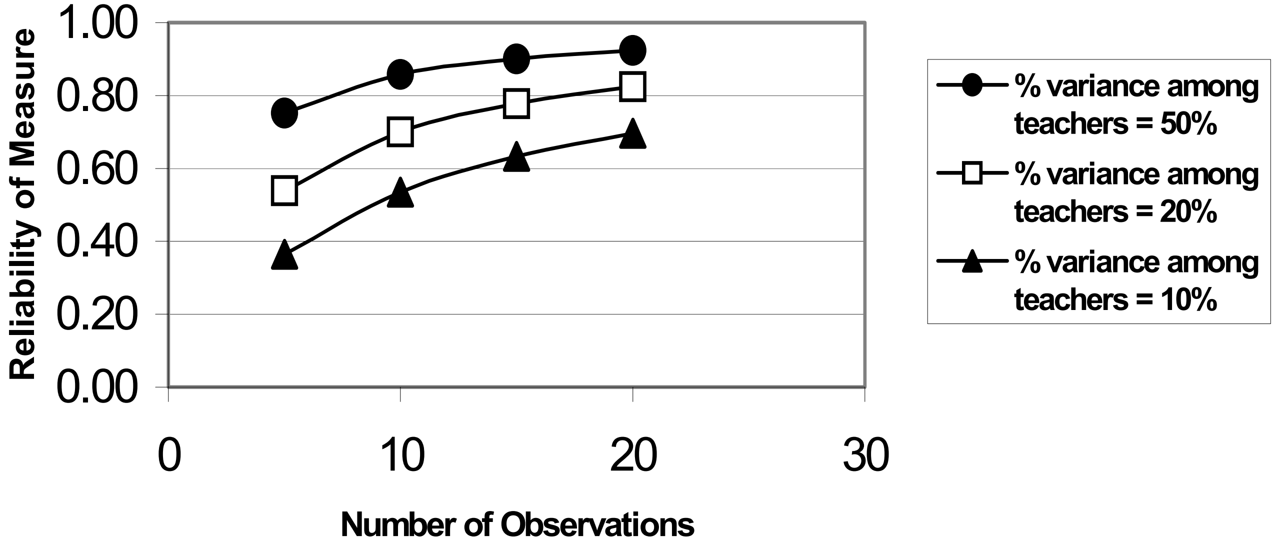
	Comprehension		Writing	
Variables	Coefficient	SE	Coefficient	SE
Intercept	-1.85	.09	-1.62	.05
Teacher:				
America's Choice	-1.52***	.26	.12	.18
Accelerated School's	-1.41***	.23	-.34*	.17
Comparison	-1.17**	.40	-.31	.28
Has Master's degree	.44*	.22	.04	.16
Missing data on Master's degree	.88*	.43	.04	.16
Amount of professional development on teaching methods	.22*	.11	.23**	.08
Years experience	.01	.01	-.00	.006
Classroom:				
Average Fall Achievement	.00	.002	.003	.003
Average SES	.11	.24	-.20	.18
Percentage of white students	-.15	.30	-.22	.23
Self contained	.37	.30	.02	.22
Missing data on self contained	1.05	.59	.41	.42
Student:				
Fall achievement	.001	.001	-.00	.001

Male	.00	.05	.04	.07
White	-.03	.10	.03	.11
SES	-.07	.05	.06	.05
Engagement (Teacher rating)	.04	.05	-.02	.05
Occasion:				
Log date	.002***	.0005	-.003***	.0005
Monday	.06	.09	-.08	.09
Tuesday	.10	.09	.07	.09
Wednesday	.07	.09	.04	.09
Thursday	-.02	.09	.18*	.09
Holiday	-.08	.11	.16	.12
Focus of Lesson:				
Comprehension	--	--	.15***	.02
Writing	.19***	.02	--	--
Word analysis	.12**	.04	-.08*	.04
Concepts of print	.19**	.05	.40***	.05
Reading fluency	.43***	.03	-.16***	.03
Vocabulary	.17***	.03	-.13***	.03
Grammar	-.07	.04	.40***	.04
Spelling	-.10	.04	.36***	.04
Research strategies	-.02	.05	.12*	.05
Summary statistics:				
Fully unconditional model:				

Variance component:		
Occasion	3.406	3.439
Student	.001	.131
Teacher	2.078	.740
Reliability:		
Student	.001	.183
Teacher	.902	.753
Prediction model:		
Variance component:		
Occasion	2.986	3.005
Student	.002	.086
Teacher	1.060	.431
% Reduction in teacher variance from table 1	49%	42%

* p<.05 ** p<.01 *** p<.001

Figure 1: Reliability of Log-Based Measures and Number of Observations Under Different Scenarios



NOTES

This article is a revised version of a paper presented at the annual meeting of the American Educational Research Association, April 2002, New Orleans, LA. Work on the article was supported by grants to the Consortium for Policy Research in Education from the Atlantic Philanthropies (USA), the National Science Foundation, and the U.S. Department of Education. We gratefully acknowledge the advice and assistance given to us at various stages in the work by Sally Atkins-Burnett, Andy Hayes, and Robert Miller.

[1] The *Study of Instructional Improvement* is being conducted by the Consortium for Policy Research in Education (Deborah L. Ball, David K. Cohen, and Brian Rowan, principal investigators). Its purpose is to examine the design, implementation, and effectiveness of three of the largest instructional improvement programs in the U.S.: the Accelerated Schools Program, America's Choice, and Success for All. Over the course of the study, data will be collected on many features of families, students, classrooms, and schools, so the study is investigating issues that range well beyond the enacted curriculum. Readers interested in learning more about the theoretical underpinnings, the research questions being investigated, and the instruments being used in this study can consult the project's web site at www.sii.soe.umich.edu.