Introduction: Measuring Instruction

Deborah Loewenberg Ball
Brian Rowan
University of Michigan

The quality of teaching in U. S. schools is of central concern to policy makers, researchers, and the public. Policy makers demand that there be a qualified teacher in every classroom; researchers examine the nature and demands of high-quality teaching; and parents expect their children to be taught by able, caring, and dedicated teachers. These concerns for teaching make sense. Research demonstrates that once children enter school, teachers exercise more influence on students' academic growth than any other single factor, including the families in which students grow up, the neighborhoods where they live, and the schools they attend (Sanders & Horn, 1994). However, research also shows that teachers vary enormously in their ability to boost children's academic growth. Rowan, Correnti, and Miller (2002), for example, showed that two students from identical social and academic backgrounds assigned to classrooms with similar student composition inside the same school can experience widely varying rates of achievement growth due to differences in their instruction. In the face of such findings, it makes sense to be concerned about the quality of teaching and to seek to ensure that all students have good teachers every year they are in school.

This special issue of the *Elementary School Journal* has its origins in this set of concerns about instruction. It is increasingly clear that instructional quality affects what students learn in school and how they grow academically over time. However, less is known about what makes teaching good, or effective. Researchers also lack adequate knowledge of how to measure good teaching, assess its effects on students' academic

The Elementary School Journal
Volume 105, Number 1
© 2004 by The University of Chicago. All rights reserved.
0013-5984/2004/10501-0001$05.00

achievement, and promote such teaching in schools. Out of interest in these issues, we and our colleagues at the Consortium for Policy Research in Education launched the Study of Instructional Improvement (SII) several years ago.[1] The SII is a large-scale, multimethod, longitudinal study of the effects of three of America's largest comprehensive school reform programs (Accelerated Schools, America's Choice, and Success for All) on instruction and student achievement in a sample of 90 high-poverty urban elementary schools. The study examines the design and operation of the school reform programs and carefully assesses the extent to which these programs affect instruction and student achievement in the core academic areas of reading/language arts and mathematics. Our design includes 30 additional comparison schools that are not involved in the reform programs. In addition, we are conducting case studies in 12 schools selected from the overall sample—nine that are participating in one of the school reform programs and three that are not.

The Measurement of Teaching

Because SII is designed to focus on the nature and quality of instruction, we have become interested in an area of research that Shavelson, Webb, and Burstein (1986) called the "measurement of teaching." In using quantitative methods to conduct systematic research on teaching in large samples of schools, we were not satisfied with the methods available to measure teaching on a broad scale. Indeed, our review of the literature on large-scale survey research on teaching suggested that many studies use inexact measures of doubtful reliability and validity (for a review of the literature in this area, see Rowan et al., 2002). As a result, in designing the Study of Instructional Improvement, we attempted to develop better measures of teaching.

The study of instruction has a substantial history in educational research. Some scholars have used infrequently administered surveys of teachers to obtain patterns of curriculum coverage and emphasis, as well as to measure instructional processes characterizing the delivery of that curriculum. Because gathering annual data on daily instruction likely often misrepresented actual practice, more frequently administered logs emerged as an approach to gathering information about content covered (Knapp & Marder, 1992; Porter, 2002; Porter, Floden, Freeman, Schmidt, & Schwille, 1986). Still other researchers have used interviews to gather information about teachers' instructional practice. One promising strategy has involved posing scenarios of teaching situations and asking teachers how they would respond in the situation (Kennedy, Ball, & McDiarmid, 1993). Yet another approach has been to collect artifacts from classrooms and to use those to analyze students' opportunities to learn, their engagement in learning, and teachers' interactions with students over instructional tasks (Borko, Stecher, Kuffner, Arnold, & Wood, 2004; Burstein et al., 1995). Classroom observation has been yet another widely used method of gathering information about instruction. Such observation has involved either detailed field notes of teachers' and students' activities, videotaping, or the use of more structured checklists or codes to reduce the data into categories of interest (e.g., who talks, content focus, nature of the activity). Some consider classroom observation the "gold standard" for collecting information about instruction. There are, however, reasons to recognize both assets and limitations in each of these approaches.

Researchers using these different methods have too often divided themselves into different camps, talking past one another and disputing one another's findings. Some researchers, for example, have concentrated their efforts on producing broad-scale views of instruction, using survey research methods to gather information on a few elements of practice across many classrooms and schools. Others have used a finer-grained lens, studying practice in detail in a few classrooms. Each approach affords useful perspectives on what is happening in classrooms; however, each also lacks insight gained by viewing instruction from other angles. Thus, broad-scale studies can lack detail. Validation is also lacking because key descriptors of practice used in survey instruments are seldom understood uniformly by respondents. Meanwhile, smaller-scale but in-depth studies can also suffer from problems of validity and reliability, for the quality of the data is so dependent on the observer. Moreover, small-scale studies lack confirmation that the patterns seen hold for the larger population. Thus, one reason that the task of measuring instruction is a challenge is that both large-scale portraits and close-up detailed information are needed in order to understand teaching.

The Study of Instructional Improvement developed an array of methods to collect data on instruction. These included annual questionnaires administered to teachers; instructional logs teachers completed frequently during the school year; classroom observations and teacher interviews, conducted over a period of years with the same teachers; and document collection and analysis. Along with these measures, we designed a variety of instruments to study the environments in which teaching occurred. These included observations, interviews, and questionnaires that assessed school culture and climate, instructional organization and management, and leadership and staff development, as well as a variety of survey instruments to learn about the social background, motivation, and social development of the students in these schools. Our goals were to produce high-quality methods for collecting information about teachers, their instruction, and the environments in which they worked. We sought, too, to create tools that would be useful to and usable by other researchers.

We have chosen in this special issue to focus on SII researchers' efforts to develop better measures of teachers' content knowledge for teaching and of the enacted curriculum in schools. Both these dimensions of quality teaching are likely to exert effects on students' achievement. Noteworthy is that our measurement work in these domains is within the tradition and mode of survey research in education. Although each of the authors in this special issue has been involved in close-up, qualitative studies of teaching and teachers' knowledge, the work we present here centers on our research group's efforts to design large-scale approaches to the study of teachers and teaching. This work challenged us. The challenge was attractive for many reasons, not the least of which was that portraits of elementary instruction and teachers in this country are spotty—detailed in some ways and extrapolated unreliably in others. We hoped that our work, demanded by our own large-scale study, could contribute to the growing need for reliable strategies for measuring instruction on a broader scale. It is that wider need, and conversations about how to better meet that need, to which we aimed the articles in this issue.

Challenges in Measuring Instruction

We turn now to a few observations about the special challenges of measuring instruction, not only in large-scale research, but generally. We identify six such challenges, explain their significance to any effort to measure instruction, and discuss how we have contended with each of these challenges in our own work. These introductory comments provide an orientation to the articles that follow.

One challenge in measuring instruction is related to sampling (Rowley, 1976, 1978; Shavelson & Dempsey-Atwood, 1976). At what frequency should teaching be documented? And what about it should be documented? Decisions about sampling depend on purposes. Because we are collecting data on instruction in order to document students' opportunities to learn, we have used instructional logs to collect information on teachers' instruction at the level of the *individual day.* Instead of asking teachers retrospectively about their instructional practice in general, or over long intervals of time, we asked teachers for records of their instruction in language arts and in mathematics on particular days. In addition, we asked teachers to report on the instruction offered to a single student in their classes, sampling across students over time. In our design, students were selected randomly from their class rosters, and across days teachers were asked to report on different students in the sample. This strategy allowed us to collect information about individual students' opportunities to learn, including the regularity with which they received instruction in either language arts or mathematics, the amount of instruction they received, and descriptive information about those opportunities on particular days.

This approach to collecting instructional data with reference to individual students on certain days was designed to specify more clearly the sampling frame within which instruction is being documented. Approaches that ask teachers to report retrospectively on instruction, over all days of instruction, and across all of the students they taught, might obscure the instruction particular students received across a year, as well as the variability in instructional practices over time (Mayer, 1999). Our approach seeks to offer a clearer picture of these elements. In contrast, our decisions about the kinds of instructional activities to record on the log represent choices about what to sample from the wide range of activity that comprises instruction. Concerned with minimizing teachers' burden, we sampled dimensions of teaching based on our own judgments about the feasibility of collecting good information on a practice or on the likelihood that a practice would affect achievement. For example, we do not ask about grouping practices used on a given day in our language arts or mathematics logs (although we do gather this information in our annual questionnaire). Instead, we ask about the kinds of representations with which students work in math, about the nature of texts in use in language arts, and about the tasks that students are asked to do.

We faced sampling problems in our work on teacher knowledge as well. Because we could not ask teachers about all aspects of content and its uses in teaching, we chose topics in both reading and mathematics that were prominent in the curriculum and therefore likely to be taught regularly. Because we were interested in exploring links between teachers' content knowledge and their students' learning, we also wrote questions about areas of content that are both vital for students' progress and known to present difficulties for students. Place value is an example of such a topic in mathematics, word analysis in reading. Our approach to measuring content knowledge focuses on knowledge as it is used in practice (Ball, Bass, & Hill, 2004), and so we sampled not only across curricular areas but also across the tasks of teaching for which teachers draw on such knowledge. For instance, teachers use knowledge of content as they interpret students' responses, select examples for instruction, and provide explanations. In sampling from among the domains of work in which teachers engage, we sought to select tasks that most teachers, independent of their approach to teaching, are likely to do frequently. For example, most teachers face the challenge of making sense of students' unexpected responses to instructional tasks. Most teachers must sort out reasonable answers from those that are incorrect. Most teachers must assess the adequacy of students' performance, and most must select effective representations. In sampling the content-intensive work that teachers do, we chose high-frequency tasks of teaching in which to develop questions that drew on teachers' content knowledge.

A second challenge in developing high-quality measures of teachers' content knowledge and teaching practice is to create measures that reliably discriminate among the objects of measurement and to measure the constructs with validity. Given our interest in measuring the enacted curriculum, and because our data allowed us to assess the curricular content taught across days, students, and teachers, we decided to explore the extent to which we could obtain reliable measures of the enacted curriculum for each of these objects of measurement. Moreover, we also wished to establish the validity of our measures of the enacted curriculum. For example, did teachers' responses to our log measures correspond to those that would have been attained had we used a third-party observer? The same issues confronted us in our attempt to develop measures of teachers' content knowledge. Could we, for example, develop measures that discriminated reliably the content knowledge held by different teachers? To do this, we had to develop questions that were easier and questions that were difficult so that we could distinguish teachers with more knowledge of mathematics or language arts from those with less knowledge of these subjects. In addition, we needed to know whether our measures reflected the knowledge intended by our questions. For example, would a teacher who answered an item correctly understand the content, or might a question be answerable without knowing the ideas involved?

A third challenge in measuring instruction centers on the combination of pluralism and debates over particular views of good teaching. This represents another sort of sampling problem. Is the aim to develop measures of "reform-oriented" teaching, or to capture the range of approaches to instruction? Although perspectives on good teaching vary widely, they remain weakly specified. Moreover, the actual connections of any approach to student achievement remain unproven. To contend with this challenge in our study, we sought to develop measures that were agnostic with respect to particular views of good teaching. Because we are studying a range of approaches to improving instruction, we wanted to design methods of documenting instruction that did not presuppose the desirability or effectiveness of certain approaches. We wanted instead to develop tools that teachers with different orientations to effective teaching could comfortably complete. In being asked about a mathematics lesson on a given day, for example, we wanted a teacher whose lesson consisted of teacher explanation followed by student practice of basic arithmetic skills to be as able to record what she did as a teacher whose lesson was characterized by students discussing and seeking to prove a conjecture about even and odd numbers.

A fourth challenge in measuring instruction is to align strategies carefully with the goals of measurement. If, for example, the goal of a study is to gather information about teachers' preferred methods of teaching, then surveys asking teachers what they would do under particular circumstances are appropriate. If, however, one wants to associate what teachers do with what their students learn, as we wished to do, then hypothetical questions about practice are less useful (Kennedy, 1999). In our study, for example, we wanted to examine variation among teachers with respect to their content knowledge for teaching as well as their practice. Moreover, we hypothesized that these two elements would be related, that they would be associated with student achievement, and that they might vary within and between schools as a function of the opportunities teachers experienced in the programs we were studying. Consequently, we developed approaches that focused on teachers as the objects of measurement and on dimensions of teachers' knowledge and practice. Had we pursued different analytic goals, however, our measurement strategies might have differed.

A fifth challenge in developing measures of instruction is that such measures rely on language to make distinctions in a realm where consensual understanding is usually lacking (Hill, in press). Language is the vehicle of survey measurement, but the language of questionnaires often lacks the precision needed for reliable and valid measurement. In writing questions to which teachers could respond reliably, we faced challenges of how to express clearly and comprehensibly aspects of instructional practice. Asking, for example, whether or not students engaged in problem solving, or in discussion, or in giving explanations is fraught with difficulty because these terms, ubiquitous in teachers' talk, are nonetheless not well defined. What one teacher means by problem solving can be dramatically different from another's meaning for the term. And although most teachers would say that they conducted a discussion, the actual practice captured by that commonplace term varies widely. Further, we found that terms related to the content areas of language arts and mathematics depended on content knowledge to be interpreted: What it means for a student to be asked to prove a mathematical statement, or what integers are, for example, or what phonemes or sound segmenting represent can be unclear to teachers. Our approaches to resolving this problem were twofold. One approach was to work closely with teachers through several rounds of instrument pilot testing and on this basis to revise and refine our use of language in ways that improved terms' comprehensibility. A second was to use the piloting experience to develop glossaries for the instructional log. In these glossaries we sought to make explicit the meanings intended for terms. We used examples and tried to make clearer which items referred to which practices or ideas that teachers might wish to report about their practice. For example, on the glossary for the mathematics log, we clarified the difference between asking a student to explain how she did a problem, asking her to verify an answer, and asking her to prove that a method or a claim works in general.

A final challenge in measuring instruction is drawing on, and using wisely, the multiple sources of knowledge about instruction contained in the field. There are many sources of knowledge from which one can build in measuring instructional practice or teachers' knowledge, including research on learning; research on teaching; expert opinion (e.g., mathematicians, experts in the reading field); accomplished teachers' wisdom of practice; curriculum materials and frameworks; contemporary visions of good practice. Again, decisions about how to exploit these resources depend on the goals of the research. Because we wanted to develop tools that were usable by teachers and likely would help us investigate relations between teachers' practice and their students' achievement, we used multiple sources as we developed the domain maps and the instruments for our work. We used prior research on teaching to select aspects of instructional practice for which evidence existed of connections to students' learning (e.g., content covered) and to eliminate those where such evidence remains lacking (e.g., behavioral setting). We also drew extensively on prior research on teachers' knowledge to identify key content-knowledge issues related to teaching practice (e.g., use of representations, diagnosis/recognition of student difficulties, decompression of accomplished reading or mathematics practice into the elements that make it learnable). We also systematically sought and used expert views and critiques across many cycles of instrument design and redesign. Practice itself, in the form of repeated cycles of field tests and interviews, served as another crucial resource to design measures of instruction. We designed tools based on domain maps, drafted instruments, pilot tested them in classrooms and with teachers, sought feedback from teachers, examined data and patterns in results, and revised. Several cycles for both log and teacher knowledge measures were conducted before we completed the instruments.

In This Issue

This special issue contains five articles that report on different facets of SII researchers' efforts to measure instruction. The first two articles focus on our efforts to measure teachers' knowledge of content for teaching, in reading and in mathematics. Knowledge is one of the important broad resources for teaching. In the first of the two articles, Heather Hill, Stephen Schilling, and Deborah Ball describe both the theoretical foundations of our approach to the measurement of teacher content knowledge, as well as what has been involved in developing reliable and valid measures. The article also includes findings about the scales we have developed so far and the psychometric tools involved in building measures in this domain. The second article on teachers' content knowledge, coauthored by Geoffrey Phelps and Stephen Schilling, explores the issues involved in measuring reading, a subject in which the history of efforts to conceive and measure teachers' knowledge is much more recent.

The next three articles in this special issue shift attention from teachers' knowledge for teaching to what teachers actually do. These focus on the instructional log, how it captures instruction, and some initial analyses of what it can reveal about teaching in the schools in our study. One article, by Eric Camburn and Carol Barnes, presents a detailed account of how the research team validated the language arts log and presents results of what we learned midpoint in our development of that log that ended up informing and improving our continued development work. This article offers those interested in the problem of validation a well-developed case of the work involved in such validation. It also offers a perspective on the challenges of developing valid measures of instruction in language arts. The next two articles, one coauthored by Brian Rowan, Eric Camburn, and Richard Correnti, and the second by Rowan, this time with Delena Harrison and Andrew Hayes, report on language arts and mathematics instruction, respectively, in the urban schools that are the sites of our study. Taken together, these articles provide a view of instructional practice in elementary school classrooms that is more nuanced and detailed than is often the case in research on teaching, and one that is drawn from a larger sample of classrooms than is common. But these articles also offer a close-up look at what the instructional log makes possible and what its limitations are. As a tool for measuring instruction, the log provides answers to some important questions about students' opportunities to learn but leaves others unanswered.

In summary, this special issue is intended to contribute resources to the important problems facing those who seek to measure instruction for a variety of purposes. Whereas our work focuses on studying teaching and teachers in an effort to learn what makes instruction effective, others are concerned more with evaluating instruction or collecting broad indicators of teaching practice and descriptors of teachers' qualifications. Defining differences in purpose and method can help make clearer the issues in the measurement of instruction and make comparisons among approaches more nuanced. Finally, the special issue can help researchers focus on directions for further development of tools to measure instruction and mediate the challenges posed by this task.

Note

[1] Co-principal investigators of the Study of Instructional Improvement are David K. Cohen, Brian Rowan, and Deborah Loewenberg Ball. Assistant directors of the study are Eric Camburn, Carol Barnes, and Ruben Carriedo. The study is conducted under the auspices of the Consortium for Policy Research in Education and funded by grants from the U.S. Department of Education (grant OERI-R308A60003), the National Science Foundation's Interagency Educational Research Initiative (grants REC 9979863 and REC 0129421), the U.S. Department of Edu-

cation—to the Center for the Study of Teaching Policy (grant R308B70003), the William and Flora Hewlett Foundation, and the Atlantic Philanthropies. Opinions expressed here are ours and do not reflect the views of the funders.

References

Ball, D. L., Bass, H., & Hill, H. (2004, July). *Knowing and using mathematical knowledge in teaching: Learning what matters.* Invited paper presented at the Southern African Association for Research in Mathematics, Science, and Technology Education, Capetown, South Africa.

Borko, H., Stecher, B., Kuffner, K., Arnold, S., & Wood, A. (2004, January). *Using classroom artifacts to measure instructional practice in middle school mathematics: A two-state field test.* Presentation at an invitational conference, The Measurement of Instruction: Technical Challenges and Implications for Research, Policy, and Practice, Washington, DC.

Burstein, L., McDonnell, L., Van Winkle, J., Ormseth, T., Mirocha, J., & Guiton, G. (1995). *Validating national curriculum indicators.* Santa Monica, CA: RAND.

Hill, H. C. (in press). Content across communities. *Educational Policy.*

Kennedy, M. M. (1999). Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis,* **21**(4), 345–363.

Kennedy, M., Ball, D., & McDiarmid, D. (1993). *A study package for examining and tracking changes in teachers' knowledge* (Tech. Series 93-1). East Lansing: Michigan State University, National Center for Research on Teacher Education.

Knapp, M., & Marder, C. (1992). *Academic challenge for the children of poverty: Vol.* **2.** *Study design and technical notes.* Menlo Park, CA: SRI International.

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis,* **21**(1), 29–45.

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

Porter, A. C., Floden, R. E., Freeman, D. J., Schmidt, W. H., & Schwille, J. R. (1986). *Content determinants* (Research Series 79). East Lansing: Michigan State University.

Rowan, B., Correnti, R., & 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.

Rowley, G. L. (1976). The reliability of observational measures. *American Educational Research Journal,* **13**(1), 51–59.

Rowley, G. L. (1978). The relationship of reliability in classroom research to the amount of observation: An extension of the Spearman-Brown Formula. *Journal of Educational Measurement,* **15**(3), 165–180.

Sanders, W., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education,* **8,** 299–311.

Shavelson, R. J., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research,* **46**(4), 553–611.

Shavelson, R. J., Webb, N. M., & Burstein, L. (1986). Measurement of teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3d ed., pp. 50–91). New York: Macmillan.