Grading the Nation's Report Card

Research from the Evaluation of NAEP

Committee on the Evaluation of National and State Assessments of Educational Progress

Nambury S. Raju, James W. Pellegrino, Meryl W. Bertenthal, Karen J. Mitchell, and Lee R. Jones, *editors*

Board on Testing and Assessment

Commission on Behavioral and Social Sciences and Education

National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C.

NATIONAL ACADEMY PRESS • 2101 Constitution Avenue, NW • Washington, DC 20418

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by Award No. EA95083001 between the National Academy of Sciences and the U.S. Department of Education. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for this project.

Suggested citation: National Research Council (2000) *Grading the Nation's Report Card: Research from the Evaluation of NAEP*. Committee on the Evaluation of National and State Assessments of Educational Progress. Nambury S. Raju, James W. Pellegrino, Meryl W. Bertenthal, Karen J. Mitchell, and Lee R. Jones, editors. Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Library of Congress Cataloging-in-Publication Data

Grading the nation's report card : research from the evaluation of NAEP
/ Nambury S. Raju ... [et al.], editors ; Committee on the Evaluation of
National and State Assessments of Educational Progress, Board on Testing
and Assessment, Commission on Behavioral and Social Sciences and
Education, National Research Council.
p. cm.
Includes bibliographical references.
ISBN 0-309-06844-4 (pbk.)
1. National Assessment of Educational Progress (Project)--Evaluation.
2. Educational tests and measurements--United States. I. Raju, Nambury
S. II. National Research Council (U.S.). Committee on Evaluation of
National and State Assessments of Educational Progress.
LB3051 .G66688 2000
371.26'0973--dc21

00-008316

Additional copies of this report are available from:
National Academy Press
2101 Constitution Avenue, NW
Washington, DC 20418
Call 800-624-6242 or 202-334-3313 (in the Washington metropolitan area)

This report is also available online at **http://www.nap.edu**

Printed in the United States of America

Copyright 2000 by the National Academy of Sciences. All rights reserved.

11

Developing Classroom Process Data for the Improvement of Teaching

James W. Stigler and Michelle Perry

Of the many factors that determine student academic achievement, classroom instruction is but one. Yet it is surely an important one. Indeed, all attempts to improve education must of necessity at some point be mediated through the classroom. This is obvious because classroom practice represents the most direct means for affecting student outcomes. However, there has been surprisingly little research on this link in the chain in affecting student outcomes.

As a nation, we collect very little data on what happens inside classrooms. As Mandel (1996:3-29) wrote, "The national conversation about teaching has always been compromised by a dearth of information about the quality of practice and practitioners. . . . When dismal or promising results about student performance are reported, a new chain reaction of suppositions is often set off about the degree to which teachers are to be blamed or praised. But these suppositions are just that—hypotheses disconnected from much of a factual base that might shed some light on what is occurring, including the extent to which the observed results can be accurately attributed to teacher actions." This relative dearth of data can be blamed, at least in part, on what Burstein et al. (1995) point out as the inherent difficulty in measuring instructional practice.

Despite this inherent difficulty, we argue that the merits of these data outweigh the obstacles in collecting them. As an example of the importance of these data, here it is argued that we cannot know which instructional strategies lead to positive learning outcomes unless we know which instructional practices are being used and we cannot know which are being used without somehow looking directly at educational practices. In other words, achievement data may tell us a lot, but those data cannot tell us what should be done differently inside the

classroom. We argue that for test data to be most informative, classroom processes need to be examined. If change in student learning outcomes is observed in the tests, we still need to know whether change is due to something going on in the classroom or something independent of that.

In this paper, we make the assumption that classroom process data, especially when collected in conjunction with student achievement data, can play a critical role in efforts to improve education. We further assume, however, that such data will not necessarily improve education and that it is therefore extremely important to have an explicit idea of exactly how data will be used to improve education and by whom. In particular, we argue that researchers, policy makers, and teachers need different kinds of data and will use data in different ways to improve the quality of teaching and learning in classrooms.

Five questions guide this paper: (1) What is the nature of classroom instruction, and what implications does this have for developing indicators of instructional quality? (2) What kind of data can be collected, and what are the advantages and disadvantages of each? (3) What kind of data ought to be collected, and how will the data be used to improve the quality of instruction? (4) What are the costs of collecting data of various kinds? (5) How can new kinds of data collection be integrated into the existing National Assessment of Educational Progress (NAEP) program?

Given these issues and questions, the goal of this paper is to consider what sorts of data can be collected on classroom processes. With this goal in mind, we examine the kinds of data that are currently collected on classroom processes and evaluate what can and cannot be learned from these data. We then look beyond current research practices and make suggestions for future data collection on classroom processes.

STUDYING CLASSROOM PROCESSES

Nature of the Classroom

Having established a broad interest in collecting data on classroom processes, we consider what kind of data might be collected. Before launching into a discussion of specific data collection techniques, we need to ponder the nature of classroom instruction. The data collected and measures constructed are only indicators. To assess the validity of these indicators, we must first think through the nature of what it is they are intended to be indicators of. Indeed, a framework for thinking about the constructs that define classroom instruction provides a necessary theoretical context in which indicators can be interpreted.

Classroom instruction, first and foremost, is a complex, dynamic, goal-directed system. One goal of the system is student learning, although there certainly are other goals as well. For purposes of this paper we will assume that achievement, as measured by the NAEP, is an important overall goal of the

system we describe. The system consists of several important elements, including a teacher, students, curriculum, and materials. These elements interact with each other in complex ways. Teachers orchestrate the sequence of activities that comprise the classroom lesson. These activities represent organized behavioral interactions between students, teachers, and curriculum/materials. In addition, these lesson elements interact with key contextual factors that impinge on the classroom.

To say that the classroom is a system implies that it is more than the sum of individual features or independent dimensions. Although features might be measured to indicate indirectly the functioning of the system, it is difficult to imagine features of instruction that are always good or dimensions on which lessons should be uniformly high. For example, although in general it might be true that lessons in which students are cognitively challenged are better than lessons in which they are not, there are many instances in which repeated practice with less challenging tasks is appropriate and necessary for students' learning. This presents the researcher with a significant challenge. To define quality of instruction, one must do more than define a set of features; one must evaluate features of a specific lesson with reference to how they function in the context of a goal-directed system. Indeed, one must describe the system itself to understand the meaning of indicators.

An example will serve to illustrate the practical implications of this point. In the process-product research of the 1970s and 1980s, it was demonstrated, across many studies, that student learning of mathematics was significantly associated with rapid coverage of a large number of problems during the lesson: the more problems the teacher led students through, and the faster the pace, the more students learned as measured by achievement tests (Leighton, 1994; Leinhardt and Putnam, 1987). As often as this effect was found, however, it turned out not to hold up in cross-cultural comparisons. Japanese students achieve in mathematics at far higher levels than U.S. students, yet Japanese teachers often are found to cover only one or two problems in a single lesson, compared with 30 to 40 in an American lesson (Stigler and Perry, 1988). Clearly, the indicator of how many problems are covered has different meanings in the context of different instructional systems. U.S. teachers were using problems for repeated practice, and clearly there is something to be gained by such practice. Japanese teachers, in contrast, were using problems as the focus of students' deep thinking and reflection. Simply knowing how many problems were covered was not enough to characterize the kind of instruction students experienced.

Another truth about classroom teaching is that it is a cultural activity (Gallimore, 1996; Stigler and Hiebert, 1997). What this means is that teaching, like other cultural activities, is constructed largely out of widely shared routines that are learned implicitly and are highly resistant to change. Although in our culture we perceive variability across teachers in their approach to teaching, cross-cultural comparison reveals that such variability may be relatively insig-

nificant compared with the large differences across cultures in the ways that teachers teach. U.S. teachers, for example, have varied ways of providing feedback to students who are working on math problems during seatwork. But these variations pale in size when we realize that virtually all U.S. teachers tell students how to solve the problem before they ask the students to solve it, whereas most times Japanese teachers do not. We tend not to notice those aspects of cultural activities that are shared, focusing instead on features that vary. But it may well be that the aspects of teaching that are widely shared in a culture are the ones that have the most impact on student learning.

One important implication of this fact about teaching is that it shifts our focus somewhat from the study of teachers to the study of teaching. Because the literature on classroom indicators has been largely an American one, it has tended to focus on aspects of teaching that vary in our culture. But we need to focus as well on identifying the shared cultural scripts that underlie most or all of what we see inside American classrooms. The improvement of teaching over time may be much greater if we focus on changing widely shared scripts than if we focus on understanding variations in the competence with which teachers use the scripts.

Research Questions

Viewing classroom instruction as a complex system and as a cultural activity leads us to identify several important research questions to guide our inquiry into instructional quality.

- What kinds of instructional systems can we identify? How can we describe these systems? This will involve, minimally, identifying the key elements of the classroom lesson and describing the ways in which these elements interact.
- What kinds of quantitative indicators can we develop to assess the functioning of different types of instructional systems? What are the processes that affect these indicators? We must quantify the descriptions developed in response to the first research question if we are going to validate them across large numbers of classrooms.
- What is the role of the student in different instructional systems? What are the processes by which students learn from classroom instruction, and what characteristics of different instructional systems affect how much students learn? These are key questions, as our interest in instruction rests on the assumption that student learning is affected by instruction.
- What is the role of the teacher in different instructional systems? How can teaching be improved? Again, we assume that teachers play a critical role in shaping the nature and quality of instruction in the classroom.

Each of these general research questions can be approached through various analytic frames. For example, classroom lessons can be described on a more macrolevel in terms of activity structures (e.g., classwork or seatwork) or from a more microanalytic level (e.g., detailed analysis of discourse patterns as they unfold throughout the lesson).

Units and Methods of Analysis

Starting with the assumption that classroom instruction is a complex cultural system, we have proposed a broad set of research questions. The complexity of instruction also has implications for the units and methods of analysis we choose.

Classrooms must be studied using units that make sense and that preserve the crucial aspects of the system. These units might be relatively large (e.g., units, grade levels), but they are probably not smaller than the classroom lesson. Classroom lessons have ecological validity from the teacher's point of view. Teachers plan their days in terms of lessons: "First we'll do math, then social studies." Lessons are goal directed and orchestrated by the teacher. The explicit goal of the lesson might be a student learning goal, or it may simply be the completion of some series of activities. Regardless of the goal, the lesson itself can only be understood in relation to the goal. Although we can study the lesson through different lenses (e.g., we can study the nature of classroom discourse or the patterning of teacher-student interactions), we will need to collect information about the context in which the processes operate.

It is also important to note at the outset that both qualitative and quantitative analyses will be required in our efforts to understand and improve classroom learning. The first research question we listed is one that must be answered through qualitative analysis. Identifying parts of lessons and figuring out how the parts interact to produce student learning require a qualitative analysis of the instructional process. Once the process has been described, however, it is useful to develop indicators that can be used to validate and refine the descriptive model of instruction.

Not only do we need both qualitative and quantitative data, we also need a way to link the two kinds of data together. As we will see, this has been a problem with more traditional approaches to the study of classroom processes.

TRADITIONAL METHODS: SURVEYS AND NARRATIVE DESCRIPTIONS

Most commonly we have relied on surveys to collect data on classrooms. Additionally, narrative descriptions have been used as a method of collecting classroom data. In this section, we review those methods. In particular, first we provide *descriptions and overviews* of the data forms. Next, we examine *what we typically learn* from data collected with each of these methods. Finally, we offer an *evaluation* of each of these methods, with some attention to both the limita-

tions that each method has in terms of producing data on classroom processes and the potential of providing new insights about teaching and learning.

Survey Methods

Descriptions and Overviews

Surveys represent relatively straightforward ways to collect data on a host of issues related to classroom processes; however, surveys can take several different forms. For example, even if we are just surveying teachers, teachers can be surveyed about their recollections or their opinions with questionnaires (whose answers can take the form of a rating scale, forced multiple-choice responses, or open-ended answers), interviews, or diaries. In this section, we also include observational checklists, which in some ways resemble the other data forms in this section but in other ways resemble narrative observational records. In the remainder of the section, we provide a general description of the various types of survey methods.

Questionnaires and rating scales Questionnaires and rating scales are often used to tap classroom processes. Questionnaires and rating scales used for these purposes typically request information from teachers about the activities taking place in their classrooms. Others, including classroom observers and students, also may participate in completing questionnaires about classroom processes.

This data source can provide information about what is taught, how the teaching takes place, and how much time is spent on various topics and activities. As an example, Burstein et al. (1995:xiii) asked teachers to judge the percentage of class time spent instructing with various strategies (e.g., whole-class instruction, administering tests, performing administrative tasks). One of their major findings was that, "although the picture of teaching that can be drawn from survey data is quite general, it is probably valid, because . . . data clearly show that there is little variation in teachers' instructional strategies. The majority of teachers use a few instructional approaches and use them often." With these methods we can obtain data from a large number of informants who have direct access to the information we find of interest.

Diaries We use the term *diary* to represent teachers' records of their lessons, including lesson plans, outcomes, and the like. Diaries have been used, relatively successfully, to measure curriculum content. Given that we are concerned with classroom processes, one might wonder why we specified that diaries are used to measure curriculum content. The reason is that curriculum content has often served as a proxy for classroom practices, although it is not itself a direct measure of classroom practices. Barr and Dreeben (1983:107) defined content coverage (also commonly referred to as *instructional pace*) as the amount of curricular

material that is covered over a period of time. They argued that although other indexes of productivity *designed for judging the effectiveness of instruction* are possible, "we have selected this one because, when treated at the level of individual children, it represents an instructional condition integrally connected with learning." Another reason for focusing on diaries to measure content when we are concerned about the relationship between teaching and learning, according to Brophy and Good (1986:360) is that "the most consistently replicated findings link achievement to the quantity and pacing of instruction."

As an example, Perry (1988) surveyed nine fourth-grade teachers' mathematics lesson plan books over the course of one year and recorded which problems were assigned. She then coded each problem as belonging to one of several mathematical topics. She also measured the students' mathematics problem-solving performance, both at the beginning and the end of the school year. Problems that most children solved incorrectly at the beginning of the year were designated as representing difficult topics, and problems that most children solved correctly were designated as representing easy topics. Generally, Perry found that problem assignment was related to student learning; more specifically, she found that spending a great deal of time on a few difficult problems led to better student achievement than covering many problems, especially problems that most students could solve before receiving instruction. In this study, a diary of what instruction consisted of was used to make inferences about teaching practices that were related to learning outcomes.

Interviews Interviews, conducted face to face or by telephone, allow us to get teachers' and/or the students' views of classroom processes. We can ask what happened and we can ask for evaluations about what was reported to have happened.

Interview techniques are especially useful, compared to paper-and-pencil methods (such as questionnaires and rating scales) when the potential responses have not been determined in advance. Interviews, especially those conducted by well-trained interviewers who know what sorts of issues are of interest and which deserve lengthy commentary, are desirable when we expect complex responses because interviewers can ask respondents different questions, depending on previous answers. If the potential responses are already known, less expensive methods may be more desirable.

Checklists Checklists often have been used to document classroom processes. When using checklists, all of the behaviors of interest must be defined in advance. Additionally, observers (i.e., the ones responsible for checking off observed behaviors on a checklist) need to agree about what constitutes the observed behavior. Thus, categories must not only be defined in advance, but must also be specified as clearly as possible so that the observers check the appropriate entry.

Typically, checklists are completed by outside observers, which makes this

method different from those already discussed. In this way, checklists resemble the narrative descriptions of classroom observations, which we discuss later. However, this data form resembles the other forms of survey data in that the questions to be examined generally are already known before the data are collected.

To lay out more clearly the data that can be obtained with observational checklists, we provide a brief description of two well-known investigations that have relied on this method. As a first example, Brophy and Evertson (1976) had observers note each time a specified behavior occurred, such as teacher praise for a student's good response. From their observations and analyses, they concluded that teachers whose students had the highest achievement treated their students in a businesslike and task-oriented manner. As a second example, Stigler et al. (1987) had observers in three countries check when certain classroom behaviors and certain features of classroom organization were present. Their conclusions centered around the idea that whole-class instruction means that every student received some instruction, and teachers who relied heavily on individualized instruction had some students who, basically, were never taught. Both of these examples illustrate that checklists can provide a general snapshot of classroom life.

Uses of and Outcomes from These Methods

Survey methods are used to assess many variables related to instruction and life in classrooms. One reason these methods are used so frequently is that they are easy to use. With these methods it is easy to measure curriculum content. For example, researchers can read through teacher plan books or diaries kept for the purpose of noting what topics were covered and easily judge what was and was not taught. It is also easy to measure the amount and pace of instruction. For example, researchers can ask teachers in an interview which pages in the text were covered and can use a questionnaire to ask how much time was spent in instruction. It is also easy to measure the format of instruction. For example, researchers can ask teachers to check each form that was used on each day of instruction (lecture, small-group work, etc.).

More significantly, given the concerns motivating the present paper, these methods can even be used to measure classroom processes. For example, we can ask teachers in a questionnaire whether the questions they asked their students required short answers or reflection and abstraction; we can ask whether the students responded only to the teachers' requests or whether the students provided substantive contributions without teacher prompts. In short, researchers have used these methods successfully to document a wide array of classroom features. These methods typically have been used and analyzed in the process-product approach to classroom investigation (e.g., Brophy and Good, 1986). In general, the process-product approach assesses classroom processes—or their proxies—and relates these to student outcomes.

In addition, we note that these methods are typically used to test theories. Because survey methods must generate categories and items before the data are collected, the categories and items necessarily reflect a theoretical bias. The data collected in surveys can, for example, support or call into question a relationship that a theory would predict. In this way survey data can tell us when a theory cannot be supported and thus when a new theory is called for.

Evaluation

These methods of collecting data are used frequently, in part because they can be used on a wide scale: they are easy to administer and easy to analyze relative to other methods. The ease associated with collecting survey data makes these methods the most widely used for gathering data on classrooms. The difficulty and costliness of other methods have sometimes made them prohibitive altogether or at least have limited the number of classrooms that could be included for study (we document these more fully as these other methods are discussed).

Burstein et al. (1995:35) say that "there is still much that survey data can tell us about instructional strategy. Survey data can describe the major dimensions of classroom processes and how they vary across course levels and types of schools. National survey data, collected periodically, can document trends in teachers' use of generic instructional strategies. Such information is important for determining whether or not teaching is changing in ways consistent with the expectations of curriculum reformers and policymakers." For these reasons we imagine that the NAEP could collect and productively use these sorts of data.

Of course, with any method there are drawbacks. We see three major drawbacks to the methods just described: (1) These methods leave open many threats to validity; (2) Most significant among these threats is a lack of shared language; and (3) These methods rarely contribute to generation of new ideas and thereby do not prominently contribute to national discussion. We discuss each method in turn.

Problems of Validity Probably the most serious problem with survey methods is that responses often are not accurate, thereby making them not valid. In many instances, typical paper-and-pencil survey instruments are not to be trusted because teachers are fallible human beings and may easily forget what they have done or unwittingly skew their responses based on their individual biases. We do not mean to say that teachers are not to be trusted. What we mean is that it is sometimes difficult to produce accurate responses.

In particular, it is difficult to be precise about certain behaviors. This problem was made clear by some careful work (Mayer, 1999) on the reliability of these methods. Mayer (1999:43) writes: "We cannot rely on the individual survey questions to assess the amount of time . . . teachers use specific practices . . . because the teachers do not report their practices in a consistent manner.

Thus, the portrait of specific practices conveyed by the survey is unreliable and therefore invalid." It is much more reasonable to ask teachers what they believe than exactly what they do or how they have impacted their students with what they have done. For example, imagine how hard it would be to be precise about whether you had conveyed the concept of equivalent fractions primarily with questions, explanations, or examples. Imagine the further difficulty of knowing which of these three methods of instructional practice had the greatest positive influence on students' understanding of equivalent fractions.

Mayer (1999:43) investigated this directly by comparing teachers' responses on surveys to classroom observations of these teachers. He found that "low reliability existed for most of the practice items [i.e., items intended to measure teachers' practices] examined in this study." In short, surveys probably could never give us reliable and detailed data about classroom practice. And without reliability we cannot claim to have validly measured their behaviors.

A cousin to this problem is that those who respond to surveys are often tempted to answer questions as they imagine the researchers would like them to be answered, rather than with accuracy and honesty (e.g., Burstein et al., 1995; Cohen, 1990), thus making these methods susceptible to problems of social desirability. For example, with the recent implementation of reform-based standards, teachers are increasingly aware that their practice should reflect these standards. However, their practice may lag behind their knowledge of these standards, and so they honestly respond about what they know about the standards, even though their knowledge may not be reflected in their practice, thus making their responses on surveys inaccurate (i.e., not valid).

Although reliability is clearly a problematic aspect of relying on survey methods for documenting classroom processes, the reliability of constructs measured by surveys increases when multiple, rather than single, items are used to measure constructs (e.g., Light et al., 1990; Mayer, 1998; Shavelson et al., 1986). As Mayer (1999:43) writes: "Individual indicators of limited reliability can be grouped into a highly reliable indicator." The point here is that if we can get at a potentially important behavior with multiple approaches (e.g., use observational checklists to determine which instructional strategies were used and follow them up with interviews to learn more about how often they are used and under what conditions) or multiple items on the same measure, we are more likely to avoid problems with reliability and validity than if we rely on a single item or a single measure. Thus, we would recommend that if the NAEP were to include survey measures of teacher behavior, multiple measures should be used.

Lack of Shared Language Related to the problem of not obtaining a valid picture of classroom practices with typical paper-and-pencil survey instruments is that these instruments require an evaluation of whether teachers understand the items in the way they were intended. However, for this we need a common language that we really do not have. As Burstein et al. (1995:35) put it: "Surveys typically

cannot capture the subtle differences in how teachers define and use different techniques." For example, what one teacher means when she agrees with the item "we had a discussion" may be very different from what another teacher means when he agrees with the same item. Even something as specific as "We folded paper to demonstrate equivalent fractions" is open to multiple, potentially inconsistent interpretations (Was the paper a square or a rectangular shape to begin with? How many folds were used?), thus rendering responses invalid, even to specific descriptions.

This notion is corroborated by Palincsar and her colleagues (1998), who argue that teachers' professional development should be constructed as a "community of practice." They argue that this model deals head on with two pervasive problems in the culture of American schoolteachers: "(a) the lack of consensus regarding the goals and means of education . . . and (b) the private, personal, and individualistic nature of teaching . . . which deprives teachers of collegial and intellectual support (Little, 1992)." In other words, Palincsar et al. believe that if examples are collected and used for discussion, a common language can be developed for teaching. Besides the inherent problems associated with not having a common language when teachers respond to survey items, we note that having a common language is the first critical step toward improvement and change. In this case, a common language would enable teachers to share ideas; teachers cannot be expected to implement and evaluate new practices until this takes place.

Failure to Contribute to New Ideas Third, and perhaps most importantly, these sorts of data rarely if ever contribute to the discussion of improving practice and outcomes. Why not? Because to improve practice concrete new ideas about classroom practice are needed. Without these, we cannot expect the dialogue about classroom practice to move forward productively. And, of course, all of the methods we have discussed thus far have the questions, issues, and items defined before any data are collected, thus limiting or excluding altogether the possibility of producing new, heretofore unimagined ideas about classroom practice. In this way, survey data are much better suited to supporting or questioning existing theory than developing new theory. However, this must be qualified: when theories are not supported by data, researchers are placed in a position to refine, revise, or generate new theory. In this way, survey data have the potential to contribute to theory.

Currently, most data on classroom practice can only tell us if what we want to see in teachers' practice is there or not because people (researchers, policy makers, administrators, etc.) have predefined what *should* happen. Thus, these data can tell us what is not working but cannot help generate new ideas for improvement. To generate new ideas for improvement, we would need to obtain data that permit the development of a shared language to refer to concrete

examples of different practices. Several of the difficulties with survey methods outlined here are avoided by other methods, which we describe next.

Narrative Descriptions

Overview

Narrative descriptions of classrooms produce very different types of data, and have unique advantages and problems relative to survey data. Typically what is gained with narrative descriptions is an in-depth look at a small number of classrooms. In general, researchers who use this method send a small band of observers into classrooms. The observers then take notes—in other words, a narrative account—detailing what they see in the classrooms. The narrative notes typically are summarized and/or coded for the occurrence or absence of specified or interesting events that emerge from reading the narrative descriptions. If the observers take notes with enough detail, this method has the potential to yield multiple analyses on a variety of classroom practices.

We take work that we have conducted as an illustrative example of this method (e.g., Stigler and Perry, 1988). In this investigation we sent observers into 10 Japanese schools, 10 Chinese schools, and 20 U.S. schools. The observers or observer-trainers from the three countries met intensively before data collection began to iron out exactly what sorts of details were to be included in the narrative notes and exactly how often notes needed to be recorded (in this case, at least every minute). Then, when schools were in session, the observers took four days of narrative notes in each of two first- and two fifth-grade classrooms in each of the 40 schools in our sample. After the data were collected, they were summarized and translated into English and then coded for a variety of classroom practices that we suspected were either important across all sites and/or unique to one of the three sites. The summaries and coded notes yielded results and preliminary findings, which have since been explored more systematically (see, e.g., Stigler and Fernandez, 1995).

Uses of and Outcomes of This Method

Narrative descriptions give us a great deal of information. One of the typical uses of this method is to provide data for developing hypotheses and theories about teaching and learning. As mentioned, the narrative notes whose results were presented, in part, in Stigler and Perry's (1988) report provided opportunities for developing hypotheses, which were then tested more systematically in a controlled experiment (Fernandez, 1994). In other words, narrative descriptions are often the first step in the "descriptive-correlational-experimental cycle" suggested by Rosenshine and Furst (1973).

Evaluation

Problems of Money One of the major problems with narrative descriptions of classroom observations is that they are expensive. This is true for at least two reasons: observers need to be trained very carefully and analysis is time consuming and labor intensive.

Recall the description of how observers were trained in the Stigler and Perry (1988) investigation: observers had to be brought together from three countries. They had to work together looking at videotapes of classrooms and discussing what they saw until they could agree on what should be written down in the narrative accounts. From there, the observers who attended this international training session had to go and train the remaining observers. You can only imagine how expensive the training of the observers was for this study. (But you can also imagine how worthless the data would have been without incurring this expense!)

And then there is the analysis of narrative records. Relative to survey methods, where the questions for investigation are fairly well specified before data are collected, the questions for investigation often emerge from careful reading of the data when using methods relying on narrative observation. This makes the cost of analysis—including developing coding systems, training coders to be reliable, and so forth—very expensive.

The high cost of narrative observations means that relatively few classrooms can be included in most studies. Using only a small number of classrooms, even with very rich data on these classrooms, limits the prospect of assessing state or national practices.

Problems of Reliability We also recognize that the potential for interobserver problems is fairly high with this method. In particular, observers who take notes in the classrooms must be careful to write down, in comparable (and preferably excruciatingly precise) detail, at least everything that will later be of interest. Of course, this is not likely to happen except in researchers' and funding agencies' fantasies. Thus, researchers are left to depend on notes taken by observers who may not write down the teacher's question or may miss student responses or may neglect to note that instruction was interrupted by an announcement from the office. When this happens, the results are always limited by what was initially recorded, and conclusions must always include a cautionary note.

Even in an ideal situation, if an observational study were conducted in this fashion and found to be productive, it is hard to imagine how it could be implemented on a larger, even national scale. In particular, having a reliable group of trained observers available to collect data on a national sample seems nearly impossible.

Many of the problems associated with live observation can be overcome by

capturing actual classroom processes more precisely. In particular, video has emerged as a practical way to improve the quality of classroom data.

VIDEO RECORDS OF CLASSROOM INSTRUCTION

Video has been used for many years for the study of classroom processes. However, until recently, it was primarily used for small-scale qualitative studies, often focusing on a single teacher. Video was a natural tool for this kind of study because of its richness and because of the fact that it could be played over and over again, enabling an analyst to engage in more detailed and careful analysis than would be possible in a live observation. But the use of video does not necessarily imply a qualitative analysis. In fact, video is not a method of analysis but a means of recording ongoing activity. It consists of relatively raw records of experience. On top of video records we can build both a qualitative and a quantitative analysis, provided we collect video from a large enough sample of lessons. In fact, it turns out that video is well suited to the integration of qualitative and quantitative analyses.

The most ambitious use of video to date for research on classroom processes has been in the Third International Mathematics and Science Study (TIMSS) video study (Stigler and Hiebert, 1997). TIMSS marks the first time that videotapes have ever been collected from a national sample of teachers. In the study, national samples of eighth-grade mathematics teachers in three countries—Germany, Japan, and the United States—were videotaped teaching a complete mathematics lesson in their classrooms. The primary goal of the study was to provide national-level descriptions of classroom mathematics lessons in the three countries. A secondary goal was to ascertain the impact that policy documents such as the National Council of Teachers of Mathematics Professional Standards for Teaching Mathematics have had on classroom instruction in the United States.

Although the sample sizes were not large as far as national surveys go, ranging from 50 in Japan to 100 in Germany and 81 in the United States, they were quite large for a video study. The logistical challenges of managing such large quantities of video information are considerable. Fortunately, technological advances in the computerization of video information makes the task far easier today than it would have been just five years ago. In the next sections, we discuss some of the advantages and disadvantages of video and share some strategies we have found to be especially useful for the collection and analysis of classroom video. We especially stress strategies that help ensure the objectivity of video analysis.

Advantages of Video

Video provides a number of advantages over the more traditional methods of studying classroom processes. Unlike live observations, video greatly expands

our ability to analyze complex human interactions such as those found in classrooms. With live observations, we are limited to recording whatever an observer can record. Checklists can be useful, but it is possible for a live observer to make only a limited number of reliable judgments at the speed required for classroom research. There simply is too much going on. Video, on the other hand, can be paused, rewound, and watched again. Two observers can watch the same video independently and go back to replay and discuss those parts that they saw differently. Videos can be coded multiple times, in passes that require only limited judgments by an observer on any single pass. This makes it easier to train observers, for it is not necessary to load them up with responsibilities. Fundamentally, video gives us the luxury to take our time with the analysis.

The most important advantages of video derive from its concrete, vivid, and "preanalyzed" nature (i.e., the categories are derived from the data rather than vice versa, leaving the data open to a vast array of analyses). There are at least four major opportunities that arise because of this:

1. Video records of classroom lessons provide us the opportunity to discover ideas and alternatives not previously anticipated. Checklists and other live coding schemes imply that we know ahead of time what is likely to be seen in a classroom. Otherwise, how could we predict how to categorize it? Video allows us to go in fresh, to take advantage of serendipity whenever possible.

2. The concrete nature of video means that it is not as theory-bound as other methods of data collection. This makes the same video data usable to a far wider range of researchers than would be the case with questionnaires or live-coder observation systems. Video data, therefore, are amenable to analysis from multiple perspectives and are a natural focal point for interdisciplinary collaboration. Psychologists, anthropologists, sociologists, and others interested in understanding classroom processes can all make some use of a single video dataset.

3. Not only is video interesting to researchers from different perspectives, it has a longer shelf life than other kinds of data. Researchers of today would have little interest in reanalyzing most of the process-product data generated by classroom researchers in the 1970s and 1980s mainly because the theoretical context that motivated the collection of those data was so different from that of today. But imagine if we had videos of teaching during earlier periods of our history. These would be inherently interesting and easily appropriated by the theories of today.

4. Finally, video provides concrete referents for the words and concepts we use to describe instructional processes. In part because of the isolation of teachers, we lack a shared language for describing teaching. Certain terms (e.g., problem solving) are used frequently but rarely defined. Video images make it possible for multiple observers from multiple backgrounds to agree on the meanings of such commonly used words. Not only does this advance our scientific under-

standing of classroom processes, but also it facilitates the communication of research results to various constituencies.

Integration of Qualitative and Quantitative Methods of Analysis

Perhaps the greatest advantage of video is that it allows us to integrate qualitative and quantitative methods of analysis in a straightforward and direct way. Ethnographic researchers often work, at an analytic level, to integrate qualitative and quantitative data. But these data usually come from quite different sources—for example, participant observations and questionnaires. With video it is possible to integrate different kinds of data as applied to the same raw material, thus strengthening our understanding of each.

This point can be illustrated by describing the methods of analysis used in the TIMSS video study. In TIMSS we were able to spend a great deal of time engaged in qualitative analysis of the video images collected. As mentioned earlier, a critical objective of cross-cultural comparisons of teaching is to describe the different systems of instruction that have evolved in different cultures. There is no way to build these descriptions without the qualitative analysis that arises from simply viewing, discussing, and interpreting the video lessons. On the other hand, the descriptions constructed cannot be validated unless we can relate the descriptions to indicators that can be coded objectively from the larger corpus of videos. In our analyses, qualitative descriptions became hypotheses for objective validation. Coding procedures were defined, interrater reliability was established, and then the procedures were applied to the full set of videos.

The cycle did not necessarily stop at this point, however. Once codes had been applied, counted, and analyzed, and the results tabulated, we could go back to the videos to clarify and elaborate the meaning of the quantitative findings. For example, in coding the types of questions teachers asked students in Japanese and American classrooms, we found that Japanese teachers asked students to describe and explain more often than American teachers did. But even though questions in both countries were grouped into the "describe/explain" category, the questions seemed to differ from each other in quality. Going back to the video enabled us to see that Japanese teachers asked their students for complete descriptions of how they solved a problem, whereas U.S. teachers asked students to justify specific steps in a solution. Thus, quantitative analyses are used to validate and generalize the insights gained from qualitative study of the videos, while qualitative images provide meat and meaning for the findings obtained in quantitative results.

Software for Video Analysis

The analysis of video is notoriously laborious and time consuming, especially the kind of analysis described above, which requires investigators to reexamine

the video continually as they proceed through the analytical cycle. This fact explains why video has rarely been used on a large scale. However, such use is now more frequent, due in large part to the advent of new technologies that enable video to be encoded and stored inexpensively in digital form on a computer. Once video is digital, many tasks that were nearly impossible to accomplish using videotape can now be accomplished easily.

Digital video, in contrast to analog video, can be stored in various formats and storage mediums. Archived on CD-ROM, it is virtually indestructible and will last for 100 years or more. Stored on hard disk drives it can be served over-local area networks and the Internet, making it possible for multiple analysts to access the video from wherever they are, whenever they wish. Digital video, unlike analog video, can be played again and again without ever degrading in quality. Digital video files can be copied an unlimited number of times without any loss of quality. Most significantly, digital video can be randomly accessed, making it possible to retrieve any particular piece of video instantly.

New commercially available software exploits the power of digital video for research. One example of such software is vPrism, marketed by Digital Lava, Inc., of Los Angeles. This software is based on software developed for the TIMSS video study. The software manages large quantities of video in a multimedia database, linking the video with transcriptions, annotations, and user-definable event codes. The user interface enables the user to view video on the desktop; define a code, mark codes in the data, transcribe, and write text annotations; construct and apply time and event sampling frames, retrieving sampled video clips for coding and analysis; search and instantly retrieve video clips associated with a particular text string or event category; and export events, together with such attributes as their duration for statistical analysis. One of the most powerful aspects of this new breed of software is that it can work with video files stored anywhere on the Internet. This makes it possible for different groups of researchers, located around the world, to collaborate in the analysis of a particular set of video data. It also provides new opportunities for sharing the findings of video studies.

Problems with Video

Despite the advantages of video, and the potential of new technologies to simplify the task of organizing and analyzing large video datasets, there are issues and problems that must be kept in mind when working with video data. First, it is important to realize that video is not a veridical picture of reality, although many people wrongly assume it is. In fact, it is highly filtered and potentially quite misleading. What you see and what you do not see on video are largely determined by where the camera operator chooses to point the camera and on how wide or close he or she defines the shot. Much of what is going on in the situation being videotaped is not visible on the screen, and sometimes what is not

visible is crucial to a valid interpretation of the situation. This fact becomes quite clear as soon as one starts to analyze the contents of videotape. It is frustrating to wish the camera were pointing someplace different.

The concrete nature of video images is also problematic, even if the camera is pointed in the ideal direction. Concrete images can be quite persuasive to the human information-processing system, even if they turn out to be completely unrepresentative of what typically occurs. This fact is well known by cognitive psychologists: humans are easily misled by anecdotes, even when they are told to ignore them. There is nothing we can do about this except be aware of the potential for misinterpretation.

Another problem with video is the possibility of observer effects. Will students and teachers behave as usual with the camera present, or will we get a view that is biased in some way? Might a teacher, knowing that she is to be videotaped, prepare a special lesson just for the occasion that is unrepresentative of her normal practices?

This problem is not unique to video studies. Teachers' questionnaire responses, as well as their behavior, may be biased toward cultural norms. On the other hand, it may actually be easier to gauge the degree of bias in video studies than in questionnaire studies. Teachers who try to alter their behavior for a videotaping will likely show some evidence that this is the case. Students may look puzzled or may not be able to follow routines that are clearly new for them.

It also should be noted that changing the way a teacher teaches is not accomplished easily, as much of the literature on teacher development suggests. It is highly unlikely that teaching could be improved significantly simply by placing a camera in the classroom. On the other hand, teachers will obviously try to do an especially good job and may do some extra preparation for a lesson that is to be videotaped. We may, therefore, see a somewhat idealized version of what the teacher normally does in the classroom.

Finally, it is important to consider the issue of confidentiality in the context of video studies. Methods exist for ensuring the confidentiality of participants in questionnaire and live observational studies, but with video the challenge is far greater. How does one disguise the identity of someone who is clearly recognizable in a video? Disguising images is quite laborious. The best solution is to secure signed waivers from participants, before videos are collected, that cover the range of use-situations anticipated by the researcher. It is also possible to restrict the use of video datasets and require researchers who wish to access the data to sign nondisclosure agreements. This is not an ideal solution, however, as it means that video images cannot be used as a means of communicating study results to the public. In the TIMSS video study we produced a restricted-use dataset but also collected a few public-use tapes that could be used specifically for communicating study results to a wider audience. In future studies we plan to increase the number of videos collected for this purpose.

Practical Advice for Using Video on a Large Scale

Collect supplementary information. Although videos are notable for how much information they contain, one of the first things we notice in working with video is how much information they do not contain. Indeed, it is important to realize that video is only a partial representation of what goes on in a classroom. Often we see students working at their desks, but it is difficult to see what they are working on. Thus, the first advice we would give is to supplement a video by collecting other materials and artifacts that are relevant to the lesson. For example, student work, textbook pages, worksheets, close-up video clips of manipulatives or other materials, teachers' tests, and so on, can all be collected relatively easily. Our general rule is to collect anything that would be helpful to someone trying to understand what they see on the videotape. Teacher questionnaires and interviews also fall into this category. Often it is not possible to understand what is happening on a video without knowing what goal the teacher is trying to accomplish. Asking teachers, for example, what they intend for students to learn from a lesson is often critical for understanding a videotaped lesson.

Standardize camera techniques. It is important to note that the camera is not strictly theory free. Depending on where the videographer focuses the camera, one can get a very different view of what is happening in a classroom lesson. For this reason it is important, first, to think through what it is important to capture and, second, to standardize the procedures of camera use so that different videographers will be consistent in the decisions they make about where to point the camera. Depending on the purpose of the study, it might be necessary to use more than one camera. In any case, standard rules must be developed, and videographers must be trained to apply the rules in consistent fashion.

Clearly communicate the study's goals to the participants. When collecting video there always exists the possibility of observer bias. It is possible, perhaps even likely, that teachers will behave differently when the camera is present than when it is not. We believe that the best way to minimize observer effects is to communicate clearly to teachers what the researchers' goals are. If what you want to see is what normally happens, as opposed to, say, what a teacher could do with 20 extra hours of preparation, it is important to tell the teacher that you want to see what she would have done anyway if you had not shown up with a camera. If teachers understand the researchers' goals, our typical experience is that they will try to be cooperative.

Use intermediate representations to enhance access to video information. Video information is so complex that it taxes the information-processing capacity of the analyst. For this reason we have found it necessary to construct "intermediate representations"—representations of the content of the video that can be used by the analysts to guide their inquiry into the video.

In the TIMSS video study we used two forms of intermediate representations: a transcript and a lesson table. The transcript simply consists of a written

transcription of the talk that goes on during the classroom lesson. If the talk is in a foreign language, we also use an English translation of the transcript. We have found that sophisticated analysis of instruction requires use of a transcript. Trying to understand how lesson content unfolds in the context of verbal interchange between teacher and students is difficult without a concrete transcription of the talk.

Similarly, the lesson table provides a more content-oriented representation of the lesson as it unfolds over time. In the lesson tables constructed for the TIMSS video study, we wrote down the organization of the classroom (e.g., classwork or seatwork), the activity (e.g., teacher lecturing, class discussion), and the detailed mathematical content of the lessons as they changed through time (see, e.g., Stigler and Hiebert, 1999). There are many possible ways to make such a table; the point is that some form of table is a great help to analysts as they work to understand what is happening in the video.

Work in multiple passes. This suggestion, like the previous one, derives from the inherent complexity of video information. A single analyst cannot study all aspects of a classroom lesson during a single pass through a video, and fortunately, does not have to. Because a lesson is captured on video, it is possible and highly desirable to have analysts code the video in multiple passes. On one pass a coder can focus solely on organizational aspects of the lesson; on another, on the content of the lesson. More detailed descriptions of a lesson, such as the kinds of questions teachers ask, can be constructed on yet another pass through the tape. As we will see below, software for video analysis makes it fairly simple to integrate the results of multiple passes into a single database where layers of analysis are organized by time codes.

Use time and event sampling to increase efficiency. One mistake video users often make is to assume that they must analyze all of the video they collect. In actuality it is possible to be highly strategic, analyzing only the amount of video required to answer the question that is being asked of the data. For example, if one wants to estimate the percentage of time various instructional technologies are used in the classrooms of a nation, it is usually possible to time sample from the video. Once events have been marked in a video—for example, teacher lecture—it is also possible to event sample. The number of time slices or events that need to be examined in each lesson would, of course, depend on the frequency with which the event of interest appears in the lessons. But because the data are on video, it is always possible to go back later and increase the sample from each tape according to a preliminary analysis of findings. This is something that cannot be done in a study that uses live coders in the classroom.

Set aside some tapes to be used for code development. Analysis of videotapes is largely a post hoc process. For this reason it is important to guard against the danger of being misled by chance occurrences in a sample of video. Data miners working to discover post hoc patterns in large quantitative datasets often use part of the dataset for developing hypotheses and another part for testing the

hypotheses. This same strategy makes sense in the context of video surveys. In TIMSS a sample of nine lessons was videotaped in each country as part of a field test, prior to collecting the main study sample. These nine tapes were used for discovery and generation of hypotheses. Only after hypotheses had been generated and coding procedures developed on the nine field test tapes were procedures then applied to the full sample of tapes.

Cost and Feasibility

Many people assume that video is far more expensive than the more traditional methods, thus making it not feasible for use on a large scale. In fact, however, the picture is not so clear in this regard. In general, more traditional methods cost more on the "front end," meaning that it takes more planning, design, and training to get them into the field. Video, in contrast, is more costly on the "back end." The costs for collecting video data have dropped markedly over the past 10 years. Camcorders are cheap, as is videotape. And training for camera operators is far less exacting than the training of live observers who must achieve high levels of interrater reliability before they are sent into the field.

The real cost of video is in the analysis phase. Depending on how much analysis is done, the cost can be huge. Just the transcription of video, which we believe is generally required for analysis of the data, can cost several hundred dollars for a lesson. The cost of video analysis makes video especially suitable for some applications, but not others. For example, if what we want is an aggregate-level picture of what is happening in a group of classrooms (e.g., a nation, state, or district), it may well be worth the cost of analysis. If, however, we need a picture of teaching that is reliable at the individual teacher level, it probably will be too expensive.

In general, video data and the more traditional kinds of data can both play an important role in a portfolio of classroom process data. Video data should be used for theory generation, for validating the less expensive methods, and for the discovery of alternative instructional systems. Survey methods should be used for testing hypotheses generated through video analysis and for any study that requires very large samples of classrooms.

Conclusion

In weighing the advantages and disadvantages of different kinds of data, we must know how the data will be used to improve education. It is within the context of a use-model that we can evaluate the value of collecting any particular kind of data. In the following sections we examine use of classroom process data from the point of view of two kinds of consumers of the data: researchers/policy makers and classroom teachers. It is within these contexts that we address more specifically the kind of data to collect, how to sample, and so forth.

DATA FOR RESEARCHERS AND POLICY MAKERS

Policy makers and researchers are two very different types of professionals. They have different takes on the educational process and at times will use data differently. Still, we find considerable overlap between these two groups and similarity with respect to at least two critical features. First, both policy makers and researchers are trying to understand the educational system and, in their own ways, to effect change in this system. Second, both of these groups of professionals are working from outside the classroom. In other words, although policy makers primarily are trying to effect change and researchers primarily are trying to understand change, both are attempting to relate to the process of educational change from outside the educational arena of the classroom. Given that policy makers and researchers share these critical features, which means that the ways in which they can use data are distinctively different from the way in which teachers use data, policy makers and researchers are considered together in this section.

Researchers studying teaching and learning in classrooms are interested in such questions as: What kinds of instructional practices lead to improvements in student learning? They are interested in developing theories that link instruction and learning.

Policy makers have a slightly broader focus. Although they too are interested in links between instruction and learning (albeit not as intensively as most researchers), they are interested in issues of how policy affects changes in classroom practice much more so than researchers. In addition, policy makers are relatively more invested in communicating the results of their analysis to the public than are researchers.

In this section we take one model (Cohen and Hill, 1998) as an example to help us think about the different ways that researchers and policy makers use data. We also discuss how data could be fed back and used to inform such a model.

Cohen and Hill (1998)

Cohen and Hill (1998) conducted an investigation of teachers' adaptation to the California Framework (California State Department of Education, 1985). In their work they proposed a model for thinking about effecting change for students. Their investigation informs current debate because it addresses the relation between policy and practice.

They found that previous assumptions about links between policy and classroom practice were wrong: even with the prominence of the California Framework, very few teachers changed their practice. As Cohen and Hill (1998:41) stated, "Neither teachers' practice nor students' achievement was changed by most of the professional development that most California teachers had."

The point here is that although policy (here the California Framework) is

designed to effect change, the desired change often does not take place. To understand why this happens, Cohen and Hill developed a model. Their model has student learning, as measured by student achievement, as the ultimate dependent measure of instructional *policy*. In this model, teachers' practice is both a direct influence on student performance and an outcome measure of policy. Furthermore, teachers' opportunities to learn, and their actual learning, influence their practice and thus, indirectly through their practice, have the potential to impact student achievement. A schematic representation of this model is shown in Figure 11-1.

Cohen and Hill are not alone in their conception of the relationship of policy to student outcomes. For example, Mandel (1996:3-29) wrote, "When dismal or promising results about student performance are reported, a new chain reaction of suppositions is often set off about the degree to which teachers are to be blamed or praised. But these suppositions are just that—hypotheses disconnected from much of a factual base that might shed some light on what is occurring, including the extent to which the observed results can be accurately attributed to teacher actions." In other words, both Mandel and Cohen and Hill argue that policy rarely, if ever, has a direct effect on student outcomes, even though that is often the intent. Instead, policy has its impact through teachers' actions, and thus outcomes in student performance need to be linked to teachers' actions.

Role of Data

We see at least three different ways that data could be used to inform the model laid out by Cohen and Hill. In particular, data are needed to generate models, test models, and communicate to the public.

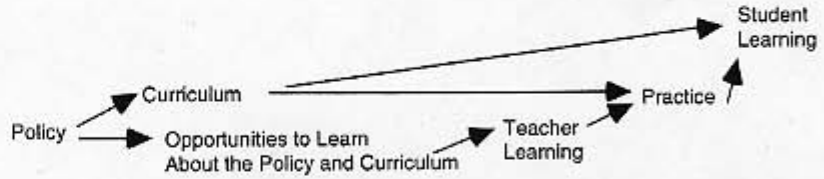

FIGURE 11-1 A schematic representation of Cohen and Hill's (1998) model of the relationship between policy and student learning.

Data Needed to Generate Models

Data can be essential for generating models—both models of the effects of policy and models of teaching and learning. Our goal here is to describe the sorts of data that allow us either to make new inferences about how policy has an impact on classrooms or to make new connections between teaching and learning. In both cases, data can be crucial for formulating useful and accurate models.

Generating Models of Policy Implementation Without models of how policy impacts classrooms, we cannot evaluate whether policy is effective. For example, even when policy is implemented, there may be problems. And without a model of how that policy ought to affect outcomes, we cannot really say why the policy did not work.

To clarify this point, we take a look at a study conducted by Cohen (1990). He reported that even when a policy makes sense and the teachers try very hard to implement it, the effect of the policy is not always felt. Cohen came to this conclusion when he found that not only did teachers adopt new practices, they also continued to use old—and in this case counterproductive—practices. Cohen voiced the concern that the adoption of new policy is not enough: pruning away ineffective, problematic practices and materials must be done if a new policy is to have its desired effect. (See also Cohen, 1995, and Siegler, 1996, for in-depth discussions of the importance of pruning away old strategies.)

One reason that Cohen and Hill suspected that the policy was not very effective was that teachers spent very little time learning how to implement policy as represented in the California State Department of Education (1985) Mathematics Framework. But Cohen and Hill also found a clear relation between the amount of time teachers spent learning about new mathematics curricula and student learning outcomes. In this case, although the policy appeared to work well, insufficient effort was expended to get the teachers to implement it.

More to the point of our concern about how data help generate models of policy and policy implementation is that collecting data on the use of a policy (i.e., implementing practices in classrooms) and not just on the ultimate outcomes (i.e., student achievement) provides insight into how to specify policy. Both are important. We note that a stated policy might easily be undermined if contradictory practices are being implemented alongside the practices recommended by the policy. Based on Cohen's work, a new model of policy implementation—one that includes adding and deleting teaching practices—should be put forth. Of course, we would not have come to this conclusion without collecting data on classroom practices. In other words, without looking at classroom practices, any picture would be incomplete of whether or how policy pertaining to classroom practice is being implemented. And in any model that links student outcomes to classroom practice, we would need to have good data on classroom practice if we are to affect (or understand) student outcomes.

What sorts of data might best contribute to the development of new models? We suspect that the most useful data in many cases would be those that take into account both the frequency and the variation in the behaviors of interest. In this case a sampling of video records would be the most productive.

Let's work through an example to make this point more clear. For example, if we believe that small-group work is important, we need to know how much time students spend in small groups, how these groups are set up (both in terms of the work to be accomplished and the composition of the groups), and, most importantly, what takes place when children work in small groups. Of course, it is difficult to imagine tracking this without a videotaped record of lessons, with the camera focused on small groups when they were operating.

Let's continue. If we find that the policy is implemented, in this case, that small-group work takes place frequently, but the effects are not as intended, we can revise the policy so that the desired outcome is more likely. For example, we may find that small groups are set up and maintained so that the students know each other well and that the students are heterogeneously grouped so that they can draw on multiple perspectives and levels of understanding. Note that these sorts of data can be obtained from survey methods. Going back to the hypothetical example, we also find that teachers almost never call the students back into a whole class and make them responsible for the work that went on in the small groups. Furthermore, when we look at what students accomplished in their small groups, we find that they spent very little time on the intended focus of the lesson. Moreover, we find that the students do not learn very much. From this scenario (which one of us witnessed over and over again), we could recommend that a policy on the use of small groups should include making students responsible for presenting their work to the whole class, thereby revealing students' misconceptions and incomplete work.

In this example, we assumed that the effects (i.e., student outcomes) were measured successfully. Of course, this should not be taken for granted. In this example of generating models that are useful for policy makers, reliable outcome measures need to be used, too. Thus, for this example to work well, we would recommend that the video records be combined with survey methods that include student achievement measures, such as the NAEP, to support the inferences that could be used by policy makers. Although it may seem obvious, it is nonetheless important to remember that when we want to know the relation between classroom practices and student performance, we need good measures of student performance. In sum, policy models can be enhanced and revised if data are collected and analyzed on implementation of the policy.

Generating Models of Teaching and Learning New ideas are often generated by data. But where do we get these data? What sorts of data would be best for generating new theories about teaching and learning? We reiterate a point made earlier: specific examples of classroom lessons are needed. If we had these

examples, a dialogue could be built about which of these lessons were good and why. In this scenario we would not have to worry as much as we do with other data sources that our language about these lessons is not understood by others: if we all watch the same lesson, for example, using folded paper to show equivalent fractions, we will know exactly how the paper is folded and how it is marked. In sum, video data can provide a shared set of examples for building language and theories for analyzing classroom practices.

Data Needed to Test and Validate Theoretical Models

The data most useful to policy makers are probably those that say whether or not teachers have implemented the stated policy and, if so, what the impact of the implementation has been on student achievement. This can then be related to student achievement data: if students perform well, the policy should remain; if students perform poorly, the policy should be revised. Thus, the first concern for policy makers is to know whether policy is being implemented. If the stated policy is indeed being implemented, it is also important to know *how* it was implemented.

Here is an example of this issue: the National Council of Teachers of Mathematics recommends that students participate in mathematical discussions. Among the many reasons for making this suggestion is that research has told us that students learn better when they participate actively than when they are passively taking in what the teacher tells them. To see whether insisting on discussions is indeed a good policy to be recommended to all teachers, we would want to know how frequently—and how well—teachers engaged their students in mathematical discussions, especially in relation to the amount of time teachers expect their students to be more passive (e.g., when the teacher stands at the front of the room and explains to the students what she wants them to know). When we know the absolute and relative amounts of time spent in mathematical discussions versus just listening to the teacher, we can relate these to student outcomes.

How do we get these data? We can imagine several scenarios, but for this sort of question we suggest that none involve teacher self-reports because teachers cannot possibly teach and note when they are using different instructional techniques and also report how much time they spent in these episodes. Thus, we recommend videotaped observations because they permit a careful and relatively accurate measure of what teachers do and do not do in their classrooms.

We also acknowledge that different types of data may be necessary to test theoretical models of teaching and learning than the types of data used to develop the models. For example, we can use videotaped records of classroom instruction to develop ideas about what might facilitate learning and then test these ideas using experimental methods. As an example, Flevares and Perry (2000) discovered that teachers vary their presentations of nonverbal information to accompany the verbal content and activities in a lesson. From this discovery, they

hypothesized that the naturally occurring nonverbal information may be crucial to learning the lesson content. At this point, Flevares and Perry (1999) are systematically presenting the same lesson content in verbal form but varying the nonverbal forms and then measuring learning outcomes. Eventually, they expect to understand which nonverbal forms aid learning of different concepts.

We also wish to make the point that even when we have what we believe is a good policy, video data can clarify the policy. This point is important because policy, such as that reflected in standards, is typically vague. When policy is vague, it leaves plenty of room for interpreting and misinterpreting. As Cohen (1990:313) puts it, "The [California] framework's mathematical exhortations were general; it offered few specifics about how teachers might respond, and left room for many different [implied: some bad] responses." Thus, we suggest that clear examples, especially those derived from videotaped observations, not only allow the development of a shared language about what practices actually reflect policy—and which do not—but can hone and clarify the policy. In sum, a wide array of data forms may be necessary to test models of the effects of policy and to test theories of teaching and learning.

Data Needed as Basis for Communicating to the Public

Finally, we raise the point that data are also needed to communicate what has been learned to the public. What sorts of data are these? Of course, the answer depends on the type of data that best illustrate what we have learned. Here is a simple example: if we have learned that teachers who spend a great deal of time learning about a new curriculum do a better job of teaching it than teachers who spend little time learning about the new curriculum, we simply need to present the average number of hours spent in training of the teachers whose students learned the material well compared to the teachers whose students did not.

Let's turn to a more complex example. If we learn that stating the goal of a lesson in a clear fashion at the beginning of a lesson facilitates students' understanding of the lesson's content, we may need demonstrations of different teachers stating the goal of their lessons. Data of this sort would allow the public to get a sense of how powerful these opening goal statements can be, especially when these are compared to other teachers' opening statements, which do not include goal statements. The general point we wish to make is that the data we share with the public need to be accessible and the data need to communicate or demonstrate clearly what can be learned.

Recommendations

Classroom process data relevant to the needs of researchers and policy makers are scant. In general we need more data of all kinds that can feed information from the classroom back into the research and policy process. Specifically,

however, we stress the need to expand our data collection efforts beyond traditional surveys. We recommend three new initiatives.

First, we desperately need to collect more data on how policies are implemented and their effectiveness inside classrooms. We need to know whether policies are implemented or not, and we need to understand the conditions under which they succeed or fail. Student outcome data must be linked into this effort, but outcome data alone will not be enough to understand how policies work. In particular, we propose that video surveys be used, in conjunction with more traditional surveys, to study classroom processes. Through questionnaires we can find out, for example, about teachers' opportunities to learn about new policies or new curricula. Through video surveys we can see what the new policy or curriculum looks like as it is implemented in classrooms. Clearly, both kinds of information are needed if we want to understand the mechanisms by which policy affects teaching and learning.

Second, apart from policy, we should conduct video studies to aid in the development of theories of teaching and to validate survey instruments. Video data are especially useful for theory generation. Recall the example we presented earlier in which we discussed "describe/explain" questions. Japanese teachers asked their students to describe complete problem solutions, whereas U.S. teachers asked students to present and justify single steps in a solution. Given that Japanese students outperform their U.S. peers, we could use this information to advance our theories of learning. In particular, we could hypothesize that it is not enough to retell one portion of a problem's solution and have others tell about other portions. Instead, for deep learning to take place, students may need to put their explanations in the context of whole-problem solutions. This hypothesis, generated from video data, could be tested experimentally. Video records also allow for validation of other instruments (see, e.g., Mayer, 1998).

Aside from general surveys, we can think of two kinds of data collection efforts that would be especially valuable. One would be the establishment of a national sample of "indicator" districts or schools that could serve as a testbed for developing theories of teaching and new survey instruments. We would propose to collect all sorts of data in these schools, including, but not limited to, achievement data, survey data (from teachers, students, parents, and administrators), and videotaped observations of lessons. In these settings, quantitative data could be linked with rich contextual data to yield important insights. Moreover, with the availability of multiple indicators and videotaped records, new theoretical ideas could be explored.

Another important use of video would be to study special classrooms: either those in which students have been shown to learn a great deal or those in which new or experimental teaching techniques are being used. Such data would not only advance our understanding of what works in classrooms but also provide guidance to teachers about what the process of changing teaching can look like. Examples of teachers who are in the process of changing allow other teachers to

see what it is like to have mixed (i.e., new and old) practices (e.g., Cohen, 1990) and can provide teachers with direct knowledge of what may be problematic in adopting something new. In addition, examples of teachers who have accomplished a successful change can provide a model, replete with explicit tactics for instructional success. Our point is simply that special cases may well be more useful than random samples in advancing our knowledge of teaching and how to improve it.

Our third recommendation is to conduct international studies in order to increase our exposure to novel variations in teaching practices. New ideas are essential if we are to improve teaching. Systems, and individuals, have a difficult time learning without a steady diet of variability (Siegler, 1996). Innovations, alternative images, different ways of doing things, and new information are all needed to create new experiences from which the system can learn (Stigler and Hiebert, 1999). Looking across cultures can be an especially useful source of new ideas about what is possible in classrooms, but only if we use research methods that can spot what is new. Questionnaires are not well suited to this goal because on them teachers can only answer the questions the researchers were clever enough to ask. Video data, especially those that are collected outside our own country, can serve this function of generating new ideas and new hypotheses about teaching.

DATA FOR CLASSROOM PRACTITIONERS

We have described the role that data can play in helping researchers and policy makers understand the chain of influence that relates policy to classroom practice to student learning. But what about classroom teachers? What role can data play, if any, in teachers' efforts to improve their own practice?

The traditional view is that teachers can use the findings from research, and the recommendations of policy makers, to improve their teaching. So, for example, teachers are assumed to read documents such as the *NCTM Professional Standards for Teaching Mathematics* and be able to use the recommendations therein as a guide for improvement. Recent data and a lot of experience suggest, however, that teaching is not easily changed by having teachers read such documents (e.g., Stigler and Hiebert, 1997). The reason, we believe, is that general research findings, because they are general, are not situated in the complexities of classroom life. As we pointed out earlier, there are few features of instruction that are always desirable or always undesirable; it depends on the lesson context.

We propose an alternative to the traditional view. Because teaching is so complex, general research findings will have limited applicability to the improvement of practice. Such findings can serve as a guide, but they will not be sufficient. Teachers need a different kind of knowledge as well, knowledge we might refer to as localized theories grounded in practice. Teachers themselves will be the ones to develop this kind of knowledge.

What Teachers Need to Know to Improve Practice

Much has been written about what teachers need to know to perform their craft (e.g., Shulman, 1986). We will not review that literature here except to point out that there is a marked difference between the kind of knowledge teachers use, as indicated by post hoc analysis, and the kind of knowledge teachers have available in their quest to become better teachers. Most attempts to improve teaching through workshops, courses, and so forth, provide knowledge that is of limited relevance in the classroom. On the one hand, teachers are exposed to theories, generated by researchers, that are decontextualized and difficult to link to classroom practice. On the other hand, teachers are given models or examples of what they "should do" in their classrooms and asked to copy them. But in these cases the examples are not grounded in theory and thus are not easily adaptable in local classroom contexts.

Our view is that teachers, to improve their practice, need a kind of knowledge that has been in short supply to this point: theories linked with examples. This is what we mean by localized theories of teaching. To be useful, such knowledge needs to be organized around curricular goals and needs to be packaged in units that are shareable across teachers and classrooms. Currently we have no means of generating this kind of knowledge, no means of accumulating and storing this knowledge, and no mechanism for sharing this knowledge across teachers. A major goal of data collection about teaching, therefore, should be to produce data that can contribute to producing theories of teaching linked with examples, and that can help in the accumulation and sharing of this knowledge.

Role of Data for Improving Teaching

We believe that teachers must play a central role in the generation of localized theories of teaching and learning in classrooms. Teachers are the ones with the best access to relevant information about classrooms, and they are in the best position to evaluate the validity of localized theories. In addition, there are many more teachers in the country than there are educational researchers. Unless teachers are involved in a central way in this process, progress will be exceedingly slow. Of course, it will take more than data to engage teachers in this process, but data can play a central role.

Generating localized theories of teaching will require prolonged reflection and discussion of examples of classroom practice. Video can play a central role in these discussions because it allows what is normally a complex and transitory phenomenon to be slowed down and replayed for study. The theoretical descriptions of teaching that can result from analysis of classroom videos will naturally be linked to actual examples of classroom practice. Thus, what teachers learn from joint analysis of such examples will be easier to situate in terms of their own classrooms. The collaboration is important, too, for it means that teachers will be

developing a shared language for describing the events and activities they see on video. This shared language is critical as it becomes the foundation on which localized theories of teaching can be stored, accessed, and communicated about with other teachers.

In the process we envision by which teachers could use classroom videos, it is interesting to ponder what kinds of examples ought to be collected. Some might think that the most important videos to analyze would be those that teachers collect in their own classrooms (see, e.g., Lampert and Ball, 1998). Although there certainly is a place for such examples in the teacher development process, they are by no means the only or even the most important examples. Because teaching is a cultural activity, and because variation in teaching methods might therefore be limited in a single culture, it is probably most important that teachers gain exposure to genuine alternatives, examples that depart significantly from what they are accustomed to seeing. Even risking possible misinterpretation, videos of lessons from other cultures, and videos of lessons in which serious efforts to reform are evident, would be a high priority for teachers because these present clear alternatives to typical and/or culture-bound lessons.

For teachers, contextual data about the lessons taped are even more critical than for researchers and policy makers. Teachers need to know what happened yesterday and what the students knew and understood before the lesson started. Test data and interview data from students both before and after a lesson would be highly relevant to teachers' analyses. Interviews with the teacher on the video would also be important, especially questions that elicit from the teacher explanations of what she or he was intending to accomplish with each part of the lesson. For teachers, the key is not sampling: lessons need not be representative, and the number of lessons need not be large. What is important is that the cases be selected to expand and inform teachers' developing understandings of teaching and learning in classrooms.

Finally, there is one more function that can be served by access to video examples. As noted by Cohen and Hill (1998), analysis of the possibilities exemplified by other teachers can provide a powerful incentive for teachers to improve their own teaching. We are reminded of the beginning Japanese teacher described by Lewis and Tsuchida (1997) who broke down in tears after watching one of her senior colleagues teach a science lesson. She explained that she thought the other teacher was so skilled that she felt badly for her own students, who, through the luck of the draw, ended up in her class. The result was a strong feeling of wanting to improve, coupled with concrete images of what improved teaching might look like.

Recommendations

Teachers can videotape themselves at the local level, but the federal government can play an important role in collecting, and then giving teachers access to,

variant examples of teaching in different cultures, different subject areas, and so forth. The federal government also can document and collect examples from teachers who are unique, either through some special talent or through participation in systematic programs of reform.

The National Center for Education Statistics also should consider accumulating examples into a national database of video cases that could be accessed by teachers over the Internet. If rules were established to control quality, it would be possible to build and maintain a database to which classroom teachers could add their own examples. Nothing would do as much as such a database to facilitate the development and sharing of curriculum-based localized theories of teaching.

VIDEO AND THE EXISTING NAEP

Having discussed new methods of studying classroom processes and having thought through how data on classroom processes might be used by different audiences to improve teaching, we return to the question of the NAEP. In particular, we wish to address the issue of how new methods, particularly video, might be used in conjunction with the existing NAEP.

The primary focus of NAEP has been on student achievement. For more than a quarter of a century, NAEP has documented national trends in what students know and are able to do in various academic subject areas. Yet there has also been a growing interest in documenting changes in the context of achievement at a national level. Student and teacher questionnaires are now included in the NAEP as a means of measuring everything from student demographics to teacher preparation, instructional practices, school policies, and out-of-school activities.

We believe that video surveys can be integrated into the NAEP framework and that they can contribute greatly to the study of instructional practices over time. Of course, it is not feasible to videotape in every classroom included in NAEP, but collecting video records of lessons in a substantial subsample of NAEP classrooms is both practical and useful. Using techniques similar to those in the TIMSS video study, videotaping in national samples of classrooms can provide the first reliable means of tracking changes in instructional practices over time. Meanwhile, before data can be accumulated on instructional trends, video surveys can provide a means of studying the classroom mediators of such variables as race and social class. For example, NAEP already provides a means of tracking racial gaps in achievement over time. But are such gaps correlated with gaps in teaching quality and instructional practices? Video records would clearly be the best means of asking such a question, especially over time.

One way to implement such an effort would be to send videographers around the country, much as was done in TIMSS. But another possibility is even more intriguing: just as the Nielsen ratings measure television viewing by placing continuous monitoring devices in a sample of homes, NAEP could place video

cameras in a sample of classrooms and conduct continuous monitoring of classroom processes. This idea is not as farfetched as it sounds. Cameras are cheap, and the technology for connecting them to the Internet also is cheap. It would not be necessary to record all of the camera images. Instead, sampling plans could be devised to get valid and reliable pictures of what goes on inside classrooms. If NAEP assessments could be administered more frequently in this subsample of classrooms—for example, three times a year—we would have the best data ever available for studying the relation of instruction and learning inside real classrooms. This idea is feasible and should be considered seriously.

Another use of video surveys in NAEP should be to aid in the development and validation of better traditional measures of classroom practices such as questionnaires. A well-designed sample of video data could serve both immediate research purposes and instrument development purposes, provided the two are integrated in their conception and design. It may be that some aspects of classroom practice are well measured by questionnaires, but validity studies to document this possibility are scant. Over time, using video in the development of questionnaires will increase the power of both methods of studying classroom practice. One way to approach this goal is to fund the development of a thesaurus of teaching practices. The problem of developing a shared language for indexing complex materials is a common one in library and information science. Library scientists have resolved the problem by relying on thesauruses, the meanings of which are painstakingly developed over time. Using similar techniques, we propose a project in which researchers, subject-matter specialists, teachers, and the public contribute to constructing a thesaurus of teaching practices linked with video examples. We believe that such a thesaurus could provide a foundation for developing new measures of instructional processes.

Yet another use of videos collected as part of NAEP would be in the communication of study results to the public. Although testing of student achievement is a complex and difficult task, the public nevertheless has some intuitive sense of what achievement tests measure. Moreover, achievement measures themselves have been validated over many years. The study of instructional practices is different on both counts. There is little agreement as to what the basic constructs are, and, as noted earlier, we lack a public vocabulary for describing teaching practices. Not only do teachers need to develop such a vocabulary if questionnaires are ever to be a useful means of studying classroom practice, but the public must do so as well if it wants to understand the information collected about classroom practices.

In terms of cost, we reiterate the fact that the cost of video data primarily resides in the analysis phase, not in the collection. For this reason we encourage the collection of larger quantities of video data, even if funds are insufficient to support in-depth analyses. Our reasoning is that an archive of nationally representative videos will become more and more valuable over time. Imagine if we had video data of instructional practices over the past 100 years. It would not be

the analyses of 100 years ago that would interest us but the opportunity for analysis now. Education is a field in which many "facts" are never really established as such, most especially those that pertain to the way things "used to be." Solid data from classrooms can play a key role in mediating and dampening the polarization that characterizes most educational debate in this country.

CONCLUSION

Data on classroom processes are critical if we are to improve education, either through policy channels, research, or teacher professional development. All attempts to improve education must, if they are to work, pass through the final common pathway that is the classroom. If we fail to collect information on what is happening in classrooms, we risk missing the key processes that could effect change. But simply collecting data is not enough. We must, before we collect any data at all, develop an understanding of how the data will be used, and by whom, to improve education. We have ruminated on how classroom process data might be used by policy makers, researchers, and classroom practitioners, but this is only the beginning. The way data are used is a subject of study in and of itself. We need more empirical studies of this process. We also need to realize that there are multiple models of data use, and so we must be flexible in collecting the data we need for different purposes.

REFERENCES

Barr, R., and R. Dreeben
1983 *How Schools Work*. Chicago: University of Chicago Press.

Brophy, J., and C. Evertson
1976 *Learning from Teaching: A Developmental Perspective*. Boston: Allyn and Bacon.

Brophy, J., and T.L. Good
1986 Teacher behavior and student achievement. In *Handbook of Research on Teaching*, M.C. Wittrock, ed. New York: MacMillan.

Burstein, L., L.M. McDonnell, J. Van Winkle, T. Ormseth, J. Mirocha, and G. Guitton
1995 *Validating National Curriculum Indicators*. Santa Monica: RAND Corp.

California State Department of Education
1985 *Mathematics Framework for California Public Schools: Kindergarten Through Grade 12*. Sacramento: California State Department of Education.

Cohen, D.K.
1990 A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis* 12:311-329.
1995 What is the system in systemic reform? *Educational Researcher* 24(9):11-17, 31.

Cohen, D.K., and H.C. Hill
1998 Instructional Policy and Classroom Performance: The Mathematics Reform in California. Paper presented at the NCTM Research Presession, April, Washington, D.C.

Fernandez, C.
1994 Students' Comprehension Processes During Mathematics Instruction. Unpublished doctoral dissertation, University of Chicago.

Flevares, L.M., and M. Perry
1999 Seeing what place value means: Building students' understanding through nonverbal representations. Poster presented at the biennial meeting of the Society for Research in Child Development, April, Albuquerque.
2000 How many do you see? The use of nonspoken representations in first-grade mathematics lessons. Manuscript under review for publication.

Gallimore, R.
1996 Classrooms are just another cultural activity. Pp. 229-250 in *Research on Classroom Ecologies: Implications for Inclusion of Children with Learning Disabilities*, D.L. Speece and B.K. Keogh, eds. Mahwah, N.J.: Lawrence Erlbaum Associates.

Lampert, M.L., and D.L. Ball
1998 *Teaching, Multimedia, and Mathematics: Investigations of Real Practice*. New York: Teachers College Press.

Leighton, M.S.
1994 Measuring Instruction: The Status of Recent Work. Unpublished manuscript, Policy Studies Associates, Inc., Washington, D C

Leinhardt, G., and R.T. Putnam
1987 The skill of learning from classroom lessons. *American Educational Research Journal* 24:372-387.

Lewis, C., and I. Tsuchida
1997 Planned educational change in Japan: The shift to student-centered elementary science. *Journal of Education Policy* 12(5):313-331.

Light, R.J., J.D. Singer, and J.B. Willett
1990 *By Design: Planning Research on Higher Education*. Cambridge, MA: Harvard University Press.

Little, J.W.
1992 Opening the black box of professional community. Pp. 157-178 in *The Changing Contexts of Teaching*, A. Lieberman, ed. Chicago: University of Chicago Press.

Mandel, D.R.
1996 Teacher education, training, and staff development: Implications for national surveys. Pp. 3-29 to 3-42 in *From Data to Information: New Directions for the National Center for Education Statistics*, G. Hoachlander, J.E. Griffith, and J.H. Ralph, eds. Washington, D.C.: U.S. Department of Education.

Mayer, D.P.
1999 Measuring instructional practice: Can policy makers trust survey data? *Educational Evaluation and Policy Analysis* 21:29-45.

Palincsar, A.S., S.J. Magnusson, N. Marano, D. Ford, and N. Brown
1998 Designing a community of practice: Principles and practices of the GIsML community. *Teaching and Teacher Education* 14(1):5-19.

Perry, M.
1988 Problem assignment and learning outcomes in nine fourth-grade mathematics classes. *Elementary School Journal* 88:413-426.

Rosenshine, B., and N. Furst
1973 The use of direct observation to study teaching. In *Second Handbook of Research on Teaching*, R.M.W. Travers, ed. Chicago: Rand McNally.

Shavelson, R.J., N.M. Webb, and L. Burstein
1986 Measurement of teaching. Pp. 50-91 in *Handbook of Research on Teaching, Third Edition*, M.C. Wittrock, ed. New York: MacMillan.

Shulman, L.S.
1986 Paradigms and research programs in the study of teaching: A contemporary perspective. Pp. 3-36 in *Handbook of Research on Teaching, Third Edition*, M.C. Wittrock, ed. New York: MacMillan.

Siegler, R.S.
1996 *Emerging Minds.* New York: Oxford University Press.
Stigler, J.W.
1996 Large-scale video surveys for the study of classroom processes. Pp. 7.1 to 7.29 in *From Data to Information: New Directions for the National Center for Education Statistics,* G. Houchlander, J.E. Griffith, and J.H. Ralph, eds. Washington, D.C.: U.S. Department of Education.
Stigler, J.W., and C. Fernandez
1995 Learning mathematics from classroom instruction: Cross-cultural and experimental perspectives. Pp. 103-130 in *Basic and Applied Perspectives on Learning, Cognition, and Development,* C.A. Nelson, ed. Mahwah, N.J.: Lawerence Erlbaum Associates.
Stigler, J.W., and J. Hiebert
1997 Understanding and improving classroom mathematics instruction: An overview of the TIMSS video study. *Phi Delta Kappan* 79(Sept.):1, 14-21.
1999 *The Teaching Gap: What Teachers Can Learn from the World's Best Educators.* New York: Free Press.
Stigler, J.W., S.Y. Lee, and H.W. Stevenson
1987 Mathematics classrooms in Japan, Taiwan, and the United States. *Child Development* 58:1272-1285.
Stigler, J.W., and M. Perry
1988 Mathematics learning in Japanese, Chinese, and American classrooms. Pp. 27-54 in *Children's Mathematics, New Directions for Child Development,* G.B. Saxe and M. Gearhart, eds. San Francisco: Jossey-Bass.