

Approximations to Indicators of Student Outcomes

Mary M. Kennedy
Michigan State University

Education policy researchers rely on a wide range of outcomes to gauge the influence of policies. Sometimes they use standardized achievement tests, and at other times they use classroom observations, teacher reports, or standardized or open-ended teacher interviews. Often, these different data collection devices are used to draw inferences about the intellectual quality of classroom events and, in turn, the quality of student learning. The purpose of this article is to examine and compare these various outcomes in terms of their potential to inform policymakers about policy influences on student outcomes.

A central problem for policy researchers is how to document a clear path of influence that extends from policy manipulations to student outcomes. There are two parts to this problem. First, because there are so many steps in the path, policy researchers may follow a policy influence only part of the way and infer that the influences they see will continue to be influential as the policy proceeds on its way to students. That is, policy researchers may examine an intermediate outcome such as teachers' attitudes toward a new policy, perhaps assuming that teacher attitudes will lead to changes in practice, which in turn will lead to changes in student outcomes. Often, however, the presumed linkages between intermediate outcomes and ultimate outcomes have not themselves been tested.

The second part of the problem is that, even when researchers seek to document influences on student learning, they are often unable to find adequate measures of the outcomes they seek. This second problem is particularly salient in the context of reforms that aim to promote more complex forms of student learning. By "complex learning," I mean learning to engage in rigorous intellectual work within a subject, drawing on important substantive concepts to solve open-ended problems, learning to reason from evidence (perhaps incomplete evidence), learning to evaluate evidence, and so forth. But this is just one definition. Terms such as *ambitious*, *conceptual*, and *authentic* have all been

brought to bear as researchers and policymakers have struggled to define the kinds of outcomes they seek. No consensus has been reached regarding the specific nature of the outcomes sought or the measurement of these outcomes. Moreover, the definitions vary across subject matter fields, so that what counts as rigorous intellectual work and what counts as a firm grasp of content will be different in history than in physics, different in English than in biology. What most agree on is that complex learning involves much more than being able to recite the main facts or rules that constitute the knowledge base in a given field.

The lack of widely available, inexpensive, and generally agreed-upon indicators of complex student learning places education policy researchers at a distinct disadvantage when examining the outcomes of various policy and program mechanisms, since many of these mechanisms aim to enhance some version of complex student learning. The lack of agreed-upon indicators further discourages researchers from trying to follow policy influences through to their influences on students and encourages them instead to rely on more easily documentable intermediate outcomes. Such intermediate outcomes are often presumed to approximate the kind of data that would have been produced if there were a suitable indicator of complex student learning.

These intermediate outcomes, or approximations, differ substantially in how well they actually ap-

proximate the elusive indicator of complex student learning. Some come relatively close, whereas others are quite removed. In some cases, we have evidence about the degree of relationship between an approximation and an indicator of complex student learning; in other cases, however, we have little or no evidence and depend instead on a hypothesized model of the path of influence that leads from policy to student learning. This article reviews a variety of approximations that have been commonly used in policy research with an eye toward defining the degree to which each kind of outcome actually approximates complex student learning. I focus in particular on studies designed to inform teaching policies, since these policies are, presumably, most likely to be aimed at enhancing student outcomes.

The article is not intended to provide a thorough review of policy research literature; rather, the goal is to provide a map of the terrain, using extant examples of research to illustrate different levels of approximation and the advantages and disadvantages of each. I raise the question of whether (and to what extent) policy research can reliably guide policy when its outcomes are only distantly related to the elusive indicator of complex student learning. If not, how can we improve the intermediate outcomes we use so that they better approximate the ideal outcome of enhancing complex student learning? I organize the intermediate outcomes according to how closely they appear to approximate this ideal, but elusive, outcome. In each case, I try to estimate how well these approximations can stand in for evidence of complex student learning.

First-Level Approximations: Classroom Observations and Standardized Tests

There are two points of view regarding where to find the closest approximation to an indicator of complex student learning. Traditionally, standardized achievement tests were presumed to indicate the full range of student learning. However, critics have begun to raise questions about the kind of intellectual work these tests require, suggesting that the range of demands they place on students is overly narrow. Consequently, some researchers who are interested in more complex kinds of student learning have turned away from standardized tests toward a very different type of approximation: classroom observation. Their assumption is that the kind of intellectual work that teachers are asking of their students might be a better indicator of the kind of work students are actually learning to do.

Thus, both standardized achievement tests and classroom observations have been used as approximations of the elusive indicator of complex student learning.

Standardized achievement tests have the advantage of ubiquitousness. They are inexpensive, convenient, and generally perceived as valid indicators of student learning. However, they have also been criticized as too narrow to use as indicators of *complex* learning. These twin features of ubiquitousness and narrowness have resulted in entire bodies of educational research being developed and then chastised for depending on standardized achievement test scores to evaluate teaching practices and policies. Two prominent examples of such work are process-product research and educational productivity research. Process-product research relies on the relationship between specific teaching practices, on one side, and gains in student achievement test scores, on the other, to draw conclusions about what counts as good teaching (see, e.g., Brophy & Good, 1986). Educational productivity studies seek relationships between school district expenditures (e.g., reducing class sizes or hiring better educated teachers) and gains in student achievement scores. Both lines have relatively long histories, and both have been quite productive. However, both lines of research are also atheoretical in that they begin with the assumption that achievement test scores are a good indicator of student outcomes and then search for correlates to these scores. Both have been criticized for this practice and for their reliance on standardized achievement tests.

The widespread availability of these tests also enables researchers to compare findings across different studies and to aggregate findings across studies. A recent example of such aggregation is Greenwald, Hedges, and Laine's (1996) meta-analysis of education productivity research. These authors used several criteria to select their studies, including a requirement that the studies' outcome measures were standardized achievement test scores. The authors' aim was to determine the extent to which additional spending to reduce class sizes, for instance, or to hire teachers with more experience or with more education, might lead to increases in student achievement. Among other things, they showed how much of a gain in student achievement test scores might be expected from spending \$500 more in any of these particular expenditure categories. But their findings do not nec-

ssarily tell us whether these expenditures would be useful for fostering more complex forms of student learning, since it is possible that different kinds of educational expenditures have different influences on different kinds of learning. Absent a model of the relationships among expenditures, standardized achievement test scores, and complex student outcomes, we cannot know how useful these productivity findings are for improving complex student learning.

As an alternative to standardized achievement tests, some researchers have turned to classroom observations to approximate an indicator of complex student learning. Observations cannot actually document the extent to which students have learned the content under discussion or the extent to which they can think through particular types of problems, but they can document the intellectual complexity of the work students are doing in class. By observing the kinds of intellectual demands that are placed on students in the classroom, we might be able to infer the kinds of intellectual work in which they are likely to show improvement.

Firestone, Mayrowetz, and Fairman's (1998) study illustrates this use of classroom observations. These researchers examined the effects of state assessment policies by observing classrooms and noting the kinds of problems students worked on in class. It is reasonable to suppose that, if students engage in a particular kind of intellectual work in school, they are likely to show improvements in that kind of work. The following is an example of the kinds of events described by Firestone et al.:

Opportunities for mathematical reasoning were often limited. In one lesson, for instance, students were given a data set on the number of sales for meals at different prices in an evening in a restaurant and asked to organize them and make inferences about how to price meals in the future. Coaching from the teacher suggested that the "correct answer" was to recommend pricing more meals in the range where the most meals had sold in the past, presumably because what had sold heavily in the past would sell a lot in the future. Alternative hypotheses...were not discussed. Thus, what could have been a real-world problem with lots of room for conjecture became highly structured and unrealistic. (1998, p. 103)

As an approximation to the elusive indicator of complex student learning, classroom observations have their own array of strengths and weaknesses. For example, observations necessarily sample rela-

tively small portions of the entire curriculum of any one teacher and sample only small proportions of all teachers, schools, or districts of interest. A number of studies conducted in the 1970s attempted to estimate the impact of teaching variability on observation findings, but most of these investigations used observation instruments not aimed at approximating complex student learning. Shavelson, Web, and Burstein's (1986) review of that literature suggests that variance can be reduced by about 20% when two observations (rather than one) are used and that variance can be reduced an additional 30% when a third observation is added. Whether these findings apply when observers are focusing on the nature of intellectual work under way in classrooms, rather than on discrete teaching behaviors, is not clear. When it comes to observing for evidence of complex student learning, we know relatively little about variations among classrooms or among days within a given teacher's classroom; thus, it is difficult to determine how representative these observations actually are or what kinds of generalizations might follow from them.

Moreover, just as is the case with direct assessments of complex student learning, there are numerous methodological approaches to classroom observation and virtually no standard practices in the field. The wide range of observation systems in use makes it hard to reconcile differences in findings from one study to the next and difficult to aggregate findings across studies. This is particularly important, because not all observations focus on aspects of instruction that might indicate complex student learning. Many observation instruments are geared toward teachers' ability to create a warm climate, for instance, or to present a lesson in an orderly sequence, or to maintain discipline. Observations such as these would not necessarily approximate an indicator of complex student learning.

If observations are to become more widespread and useful to the policy community, some standardization will be needed. Efforts by researchers such as Stigler and Herbert (1997) and Neuman, Marks, and Gamoran (1996) to define important and relevant features of classroom activities help move us in this direction.

Choosing between these two first-level approximations is a matter of deciding which aspect of the elusive indicator one most wants to approximate. Standardized tests have a cost advantage and are widely recognized and accepted, validly or not, as indicators of student learning. But they are not rec-

ognized or accepted as indicators of complex learning, and some critics contend that their widespread use actually decreases attention to more complex outcomes. Classroom observations, on the other side, are not standardized or inexpensive, no particular observation instrument is widely recognized and accepted, and those that do exist are often subject specific. And, of course, none directly assess student learning. But these devices can document the presence or absence of complex classroom learning activities that are likely to produce complex learning.

Neither observations nor standardized test scores, then, can be taken as direct indicators of complex student learning. But both are approximations that can be used, provided we have reason to believe that there is a relationship between them and the elusive indicator of complex student learning.

Relationships Between These Two Approximations and Complex Student Learning

Even though I define these two kinds of education outcomes as first-level approximations, there is still considerable dispute about their individual merits, about their relationship with each other, and about their relationship with complex student learning. Because we lack an agreed-upon indicator of complex learning, and because those indicators that do exist are not widely accepted, there does not exist a large body of research on the relationship between bona fide indicators of complex learning and either of these first-level approximations.

With respect to the relationship between standardized achievement test scores and complex student learning, Shavelson, Baxter, and Pine (1992) developed and compared a variety of indicators of complex student learning in science and included in their study a traditional standardized achievement test. These researchers found that the average correlation between students' performance on complex science tasks and their standardized achievement test scores was a moderate .43. Moreover, the correlation decreased when ability scores were partialled out. Shavelson et al. concluded that, even if disattenuated for measurement error, their complex performance tasks were assessing something different from what was assessed by their standardized achievement test.

With respect to the relationship between classroom observation data and complex student learning, we face a difficult problem, because different researchers can focus on very different aspects of

classroom life, and consequently their estimates of the relationship between what is observed and what students learn can vary considerably from one study to the next. One study that examined the relationship between observed classroom learning activities and student performance on complex tasks was that of Newmann, Marks, and Gamoran (1996). These researchers described a teaching practice called "authentic teaching" that requires that knowledge be constructed by students rather than given to them, that students develop an in-depth understanding of ideas rather than a passing acquaintance with them, and that student work have personal, aesthetic, and utilitarian value as well as academic value. Although their criteria tacitly apply to all school subjects, their research focused on mathematics and social studies classrooms. In these classrooms, Newmann et al. were able to demonstrate, by using observation studies as well as indicators of student learning, that the classroom activities they defined as authentic did indeed produce complex learning in students. Their multiple regression equation produced an R^2 value of .35. Taking the square root of that value to create a metric comparable to the simple correlations described earlier results in a correlation of .59 between learning activities observed in classrooms and their indicator of complex student learning.

Another study (Saxe, Gearhart, & Seltzer, 1999) examined the relationship between teaching practices and complex student learning of a specific topic, fractions, that constitutes a substantial portion of the upper elementary mathematics curriculum. Saxe et al. limited their analysis of teaching to those portions of class that consisted of whole-class discussions and developed two scales of teaching practice, each of which located observed practices on a 4-point scale representing a continuum from traditional, rule-bound mathematics to mathematics conforming to the standards of the National Council of Teachers of Mathematics. They used hierarchical linear modeling and did not provide correlation coefficients. However, S. Raudenbush (personal communication, July 19, 1999) helped me translate the findings of Saxe et al. into correlations. For one group of students, these authors found a correlation coefficient of .77, even higher than that found by Newmann et al. For the other group, the relationship was curvilinear. The relationships found by Newmann et al. and by Saxe et al. suggest that classroom observations that focus on the kind of intellectual work students are doing might

be better first-level approximations of the elusive indicator of complex student learning than are standardized achievement scores.

Second-Level Approximations: Situated Descriptions of Teaching

Because classroom observations are costly and perhaps unreliable as well, and because standardized tests may not approximate complex student learning very well, researchers often depend on second-level approximations rather than first-level approximations. Instead of directly observing classroom activities or directly assessing student performances, they try to obtain, from teachers, as situated a description as possible of the teachers' own teaching practices. I call these descriptions "situated" because their aim is to move past broad generalities, vagaries, or espoused principles of practice toward teachers' actual practices, but without the expense of observing these practices firsthand.

One type of situated description is the teacher log, a paper-and-pencil form, filled out by teachers either daily or weekly, that describes the details of their curriculum for a specified period of time in a specified list of courses. Logs are concrete and potentially accurate descriptions of what teachers are actually teaching in their classrooms. In addition, because they require a clear form for reporting, it is easier for researchers to convey to one another what their database consists of than is often the case with observation data.

Teacher logs were used in the Content Determinants Study (Porter, 1989; Schwille et al., 1983) to assess the extent to which external policies influence elementary teachers' enacted mathematics curricula. Figure 1 shows the teacher log used in that study. This log enabled Porter (1989) to draw a number of important conclusions about what upper elementary teachers teach about mathematics. Porter found, for instance, that teachers tended to teach for exposure rather than for mastery, that skills received about 10 times the emphasis that understanding or application received, and that the amount of mathematics instruction students receive may be doubled or halved depending on their teacher.

I consider teacher logs to be second-level approximations rather than first-level approximations to the elusive indicator of complex student learning. They are not direct observations of what was taught; instead, they are teacher reports of what was taught. If they are to be useful approximations,

teacher logs have to solve two problems. First, they must address the potential for self-serving biases, as teachers may be tempted to make their logs reflect what they meant to do rather than what they actually did. As noted by Freeman (1996), any testimony is both a representation and a presentation. Porter and colleagues' solution to this problem was to avoid value-laden language and provide content categories for teachers that were as precise and descriptively neutral as possible.

The second problem that teacher logs need to address is that teachers may have different understandings of the meanings of some of the terms included in the logs. For one teacher the cognitive task of "interpreting data" may be met by having students complete a form, while for another the same phrase may mean sorting through a disorganized collection of facts. To the extent that teachers use the terminology of teaching and learning to refer to different things, teacher logs will not reflect the nature of intellectual work done in their classrooms. The earlier event described by Firestone and his colleagues illustrates this difficulty. On paper, the problem presented to students demanded complex reasoning. But as presented by the teacher, most of the possible hypotheses had been eliminated so that the intellectual demand on students was substantially reduced. Which version would the teacher enter into the log?

There may also be a reactive quality to logs. Logs are variants of diaries, a form of data collection often used by sociologists to learn about discretionary behaviors. Irving and Elton (1986) tested the validity of diaries in a study of television viewers' use of teletext. In that study, electronic boxes attached to viewers' television sets kept records of actual teletext use. But on two occasions, 6 months apart, the researchers asked viewers to maintain diaries of their use of teletext. They found that viewers increased their use of the teletext during the weeks in which they were keeping diaries, suggesting that the very act of maintaining a diary increased their awareness of the option and, hence, their use of it.

Another form of situated description is the vignette. In vignettes, teachers are asked how they would respond to a specific hypothetical situation that is laid out for them. One advantage of vignettes is that they enable researchers to standardize situations and, thereby, aggregate findings across teachers. However, the situations are hypothetical, rather than real, so we cannot be sure

DAILY MATHEMATICS LOG			
GRID # 1: Content catalogue topics taught or assigned to S2: Student at 80 th percentile (either individually or as a member of a group.)			
TOPIC	EXAMPLES (one or two per topic)	CATALOGUE CODE	EMPHASIS (Please circle)
		(18-22)	1 2 3 (23)
		(25-29)	1 2 3 (30),
		(32-36)	1 2 3 (37)
		(39-43)	1 2 3 (44)
		(46-50)	1 2 3 (51)

Emphasis Scale: 1 = ONLY topic emphasized in lesson (emphasized 20% or more)
 2 = One of 2-4 topics emphasized in lesson
 3 = An important topic for this lesson even though not emphasized

2. Check any of the following statements that describe the lesson portrayed in Grid # 1:
 Lesson included teacher-directed group instruction
 Lesson included content-catalogue topics not covered in your primary textbook
 (note: If you don't use a textbook consider your primary source of instructional material)
 A test was given

3. Which of the following materials, if any, were used in this lesson? (Check all that apply)
 Textbook exercises, or dittos that accompany textbook series
 Other commercially prepared dittos/supplementary materials
 Teacher prepared exercises/materials (e.g., dittos or problems on board)
 Math games/puzzles

4. Among students who studied the topics in Grid # 1, did a majority work on written math assignments
 During math class? yes no
 During other periods of the school day? yes no
 At home? yes no

5. Were all three target students taught exactly the same content catalogue topics (i.e., those you have described in Grid # 1)?
 y e s ---> Stop here.
 No ----> Continue on the back of this page

FIGURE 1. Portion of the daily teacher log used in a content determinants study (second-level approximation relying on situated descriptions) (Porter; 1989).

whether teachers would actually respond to them in the way they think they would. In addition, vignettes share with teacher logs the potential for self-serving biases.

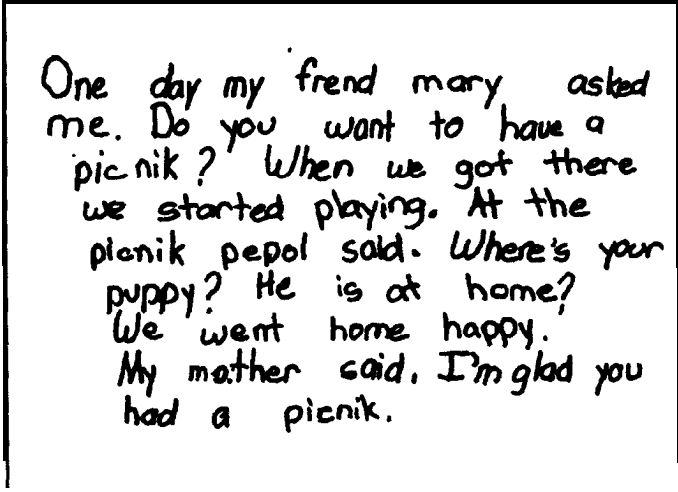
Vignettes were used in the Teacher Education and Learning to Teach Study (Kennedy, 1998) to determine whether or how teachers' ideas about teaching writing changed over time as they partici-

pated in different types of teacher education programs. Figure 2 shows an example of one of the vignettes, "Jessie's Story," used in that study.

Teachers' responses to "Jessie's Story" demonstrate that a single classroom event can evoke very different reactions. Although all of the teachers read the same story, they interpreted the story as calling for attention to different kinds of knowledge.

I would like you to imagine that your third graders are writing stories. Jessie, one of your students, hands you the following story.

- What do you think of Jessie's story?
- How would you respond to Jessie? Why?
- What grade would you assign to this paper, and why?



One day my friend mary asked me. Do you want to have a picnic? When we got there we started playing. At the picnic pepol said. Where's your puppy? He is at home? We went home happy. My mother said, I'm glad you had a picnic.

FIGURE 2. Vignette from the *Teacher Education and Learning to Teach Study* (second-level approximation relying on situated descriptions).

Kennedy defined three types of knowledge that teachers used when responding to “Jessie’s Story”: grammatical prescriptions, concepts, and strategic knowledge. The following are three excerpts that illustrate these approaches:

Grammatical prescriptions: Most likely I would have them recheck--check for grammatical errors and spelling errors to make sure that he’s written complete sentences and what have you.

Concepts: I think there are a couple of possibilities. One is that he really thinks these are different sentences. The other is that nobody has ever introduced this kid--this kid has never seen the proper notation for questions, for direct quotations.

Strategic knowledge: I might try to get him to put down more details about the picnic. I might ask, “Okay, we started playing. What did you do at the picnic?” I might ask him to close his eyes and envision the picnic. How does he feel at the picnic? Is the sun shining? What’s it like outside? I think I might do that and hopefully he will get some more details in it and that might liven it up a bit. (Kennedy, 1998, pp. 85-86)

These responses illustrate the problem of defining the content of a writing curriculum: Even though all teachers responded to the same student story, they interpreted the story as calling for very different types of subject matter knowledge. Kennedy argued that this variation in the character

of knowledge that teachers emphasized was independent of curricular topics. For instance, in a discussion of how the specific curricular topic of “transitions” in writing would be taught, one teacher indicated that knowledge of transitions consisted of rules, or prescriptions, and that “there are lots of good drills for transitions.” Another teacher thought that knowledge about transitions was conceptual: “You need a very fine sense of how the language works for the meaning that you want to get across” (Kennedy, 1998, pp. 115-116).

Ma’s (1999) comparison of Chinese and American teachers uses vignettes in a very similar way, but her vignettes have to do with teaching mathematics. For example, one vignette poses this problem:

Let’s spend some time thinking about one particular topic that you may work with when you teach subtraction with regrouping. Look at these questions:

53	91
-25	-79
-----	-----

How would you approach these problems if you were teaching second grade? What would you say pupils would need to understand or be able to do before they could start learning subtraction with regrouping?

Ma asked both Chinese and American teachers to respond to vignettes such as this one and then contrasted their responses. Like Kennedy (1998), Ma found that a single topic, such as subtraction with regrouping, could yield remarkably different ideas about what to do and about what students need to understand or be able to do before they start learning this topic. And, like Kennedy, Ma suggests that these differences reflect differences in how teachers themselves understand the situation and what kind of knowledge they think the situation calls for. For example, while American teachers talked about “borrowing,” Chinese teachers talked about “decomposing numbers.” And while American teachers talked about procedures, Chinese teachers talked about concepts.

In an early discussion of vignettes, Rossi (1979) argued that they are especially suited to learning about people’s “judgment preferences” when making choices in complex situations, which of course is what teachers are doing when they respond to

situations such as those reported by Kennedy (1998) and Ma (1999). Applied to education, vignettes can reveal teachers’ preferences for different kinds of knowledge and for different ways of representing ideas. They also have the advantage of being standardized so that researchers can aggregate data across large samples and describe patterns of variation or patterns of change in practice. In the case of the Teacher Education and Learning to Teach Study, the vignettes were used with teacher candidates who had not yet entered their own classrooms and for whom, therefore, direct observations and teacher logs were not an option. However, the same vignettes were used as candidates graduated and moved into full-time teaching positions so that changes in responses over time could be noted.

Vignettes also involve some important disadvantages. One is that it is difficult to estimate the extent to which teachers would actually behave in the ways they describe if they ever faced these teaching situations. However, as Rossi (1979) has noted, responses to vignettes tend to be lawful; that is, people’s judgment preferences form predictable patterns across situations, a fact that adds to their credibility. In addition, D. Ball (personal communication, September 1988) described an example in which she observed a teacher actually confronting the same situation the teacher had discussed in a vignette, and the teacher did in fact behave as she had said she would.

Another disadvantage of vignettes is that they are sampled from a universe of potential teaching situations, and we do not have an adequate map of that universe. If, for instance, a researcher devises an interview with 10 vignettes, we have no way of determining how well that sample represents the universe of preference judgments that are relevant to fostering complex student learning. Like classroom observations, vignettes depend heavily on the researcher’s ability to conceptualize and define the most fundamental issues associated with fostering complex student learning. The four vignettes used by Ma (1999), for instance, reflect four fundamental ideas in elementary school mathematics—subtraction, division, multiplication, and perimeter and area. Each also presents a situation that involves teaching the subject, rather than understanding the subject independent from teaching it. These vignettes, then, are useful in generating approximations to complex student learning because they focus specifically on teaching content, and the content is central to the elementary school curriculum.

Had the vignettes focused instead on, say discipline problems, or decisions about whom to call on during class discussions, they would likely be less useful as approximations to indicators of complex learning.

Finally although I have described vignettes as situated, they clearly vary in how situated they are and in which aspects are situated and which are left undefined. The degree of and location of specificity can make a difference in how teachers respond. For example, in the content determinants study referred to earlier, the researchers used vignettes to assess the degree to which teachers thought they would alter their curriculum in response to different patterns of external pressure (Floden, Porter, Schmidt, Freeman, & Schwille, 1987). Each of these vignettes suggested that the teacher had moved to a new school district where there were pressures to teach five mathematics topics that the teachers had not taught before and to remove five topics the teachers had taught before. The vignettes differed in the particular sources and combinations of external curricular pressures. In responding to these vignettes, teachers appeared to be generally compliant with external pressures, readily stating that they would change their own curriculum in response to these pressures. However, the actual topics to be added or deleted were not specified. Consequently, teachers could not consider their own personal commitment to these topics, their own experience with the topics, their feeling of comfort with teaching the topics, their private troves of materials and learning activities that they could use to teach the topics, and so forth. While the vignettes enabled the researchers to compare different patterns of external pressures, they probably overestimated the general degree of compliance teachers would demonstrate to real external curriculum pressures when particular topics that involve specific personal values are at stake. The value of a vignette, then, depends heavily on which of its features are situated and which are left unspecified.

Thus, these two approaches to situated descriptions—teacher logs and vignettes—both have advantages and disadvantages. The main advantage of the log is that it asks teachers to describe what they actually did in their classrooms, whereas vignettes do not tap into actual practice. On the other side, the log appears to be more effective in capturing topics and tasks than it is in capturing the character of intellectual work done with a topic. Logs may be more useful for subjects such as mathematics, which have relatively agreed-upon lists of top-

ics to be covered, than they would be for subjects such as language arts, where topics are less clearly defined. Vignettes make up for that weakness by focusing on the judgment preferences teachers use in specific curricular situations. In many respects, the strengths and weaknesses of these two forms of data collection complement one another, suggesting that perhaps a combination of the two might yield a strong approximation to an indicator of complex student learning. For instance, a log could indicate the extent to which teachers are covering particular topics or giving students particular kinds of problems or assignments, while a set of vignettes could be used to assess teachers' judgment preferences for, say, conjectures versus prescriptions as they manage the discussions of these topics and problems. This combination allows each form of data collection to compensate for weaknesses in the other, although both could still reflect teachers' beliefs about what they *should* do in these situations more than what they actually would do or actually did.

Relationship to First-Level Approximations

I consider the two forms of situated descriptions outlined in this section to be second-level approximations. They are not as direct as either standardized test scores or observations of students' learning activities in situ, but they are more situated and specific than other kinds of data that rely on teacher testimony. They are similar to what Koziol and Bums (1987) called *focused self-reports* in that they focus teachers' responses on specific subject matter, specific time frames, and/or specific classes or groups of students. Koziol and Bums actually tested the accuracy of focused self-reports in an observation study and found relatively high levels of agreement between focused self-reports and observed teaching practices. Agreements between teacher self-reports and observer ratings tended toward the .60s and .70s. Moreover, they rose to the high .70s and .80s when teachers used the report a second time and became more familiar with what was being asked.

Third-Level Approximations: Nonsituated Testimony About Practice

Often, researchers are unable to use either first-level or second-level approximations of the elusive indicator of complex student learning. In these cases, they rely instead on teacher questionnaires and interviews that ask about teaching practices but

are not situated in specific teaching episodes. Questions may elicit from teachers such generalities as "I always strive to promote student initiative" and "I always treat my students equally." I consider these items to be nonsituated descriptions of teaching practices. These are best thought of as revealing teachers' espoused principles of practice; they may not reveal much about teachers' theories in action (Argyris & Schon, 1996).

Obviously, there is no clear dividing line between a situated question and a nonsituated question. The distinction I am introducing here is intended to facilitate researchers' thinking about their data collection instruments rather than to prescribe clear criteria for them. Still, it does seem to me that at least two dimensions need to be attended to. First, we should aim for questions that refer to specific classroom events rather than to broad generalities. Second, we should aim for questions that define terms as precisely as possible. A recent study by Antil, Jenkins, Wayne, and Vasdasy (1998) illustrates both of these difficulties. To learn about cooperative grouping practices, these researchers first surveyed teachers and then interviewed a sample of teachers who had claimed in the survey that they regularly used cooperative groups. In the interviews, the researchers pressed for more detail about what exactly the teachers did and then compared their practices with various researchers' definitions of cooperative groups. At best, only about a quarter of the teachers were using a version of cooperative grouping that met some researchers' definitions, even though all of the interviewees had claimed in their survey responses to be using cooperative groups. No doubt these researchers' questionnaire could have yielded better data if the authors had phrased their questions more specifically and more behaviorally, even if the questions still referred to general practices. No doubt, too, it could have yielded even better data if the questions had been tied to specific classroom practices rather than to broad generalities.

Many studies of teachers' responses to new policies depend on third-level approximations. Teachers are frequently either interviewed or surveyed about their instructional practices or goals. In either case, the terms of the discussion tend to be broad generalities rather than situated descriptions. For instance, when McLaughlin (1993) sought to learn how organizational characteristics affected teachers' practices, she depended heavily on teachers' nonsituated descrip-

tions of their practice. She reported testimonials such as the following:

I will do whatever I can to get their grades up, to have them feel better about themselves . . . sometimes to the point that I think I am rescuing them instead of enabling them. (p. 86)

It is a question of pragmatism. You're doing what you can with the clientele you have. I spend perhaps an entire semester using a book that I used to open in a sophomore class for maybe a couple or three weeks. And now it's a semester. You just can't do much more. (p. 86)

We have a super-competent faculty here, and everyone who works with us says that. And there's a can-do sense. We can fix this problem. Let's roll up our sleeves and figure it out and go to work on it. (p. 91)

An example of a survey using nonsituated questions about teaching practices is Bidwell, Frank, and Quiroz's (1997) study of the relationship between forms of workplace control and teaching styles. These authors developed a theoretical model defining how workplace controls might affect teaching styles and then developed survey instruments to test their model. The survey asked about teachers' perceptions of their school and about teachers' general teaching practices. Bidwell et al. were interested in sorting teachers into different types and developed questionnaire items that they thought exemplified the practices of each type. Examples of the questions used in the Bidwell et al. study are provided in Figure 3.

Note that although these questions address teaching practices, they lack the specificity of situated descriptions. I consider such nonsituated questions about teaching practice to be more distant approximations than situated questions largely because they appear to be tapping into espoused theories of practice, and there is some evidence that people's espoused theories differ from their theories in use (Argyris & Schon, 1996). This is not to say that there may not be times when we are interested in teachers' espoused theories. In fact, Bidwell et al. did find the relationships they had predicted between organizational controls and teachers' espoused teaching styles. However, we still need an inferential leap to argue that these espoused statements actually reflect real teaching practices and that, in turn, they can be taken as approximations to indicators of student learning.

One problem with statements of espoused be-

Sample Item indicating a Progressive approach

I encourage my students to express opinions different from my own
 My assignments require students to gather information on their own
 I teach students how to learn
 My homework assignments require students to think in new ways about what I have presented in class

Sample Items indicating a Moralist approach

Order and discipline come first in my classroom
 I try to instill a common standard of values in my classroom
 My lessons are based on an explicit set of values
 I require quiet in my classroom
 Students see me as someone they can look up to

Sample Items indicating a "Pal" approach

Students talk to me about their friendships
 Students talk to me about what they do outside school
 Students see me as a friend
 Students know what I do out of school

Sample Items indicating a Rigorist approach

So far as misbehavior is concerned, I rarely make exceptions for special cases
 So far as missed examinations are concerned, I rarely make exceptions for special cases
 I refuse to negotiate with students about homework assignments

FIGURE 3. Sample of survey questions from the Bidwell et al. (1997) study of teacher types (third-level approximation relying on espoused principles of practice).

iefs or practices is that they are even more susceptible to self-serving biases than are situated descriptions. Because they are phrased in more general terms, and because they may also depend more on value-laden terms, meanings are more likely to vary from person to person. Moreover, if a statement asks about a general practice, teachers decide not only what the terms in the sentence mean but also how often these events have to occur in order for them to consider the practice to be something they "generally" do. If the statement is a socially desirable one, teachers are likely to overestimate how frequently it applies to them.

Relationship to Closer Approximations

Not many researchers have attempted to associate espoused beliefs or practices with observed classroom practices or with situated descriptions of practice. One early study, however, did: Oliver (1953) asked teachers to respond to a questionnaire listing 50 educational beliefs that he thought reflected findings from educational research. These tended to be statements consistent with progres-

sive approaches to teaching and reflected Oliver's understanding of the findings from research. He then observed the teachers' classrooms and developed a scale to define the degree to which observed practices reflected these same ideas. Oliver found that the correlation between these two measures--espoused practices and observed practices--was only .31. Moreover, average observed practice was substantially different from that implied by teachers' responses to the questionnaire. That is, teachers' espoused practices were, on average, more progressive than their observed practices. The problem of the distance of an approximation can now begin to be seen; if the correlation between espoused practices and observed practices is only .31, and if the correlation between observed practices and complex learning is in the .59-.77 range, then there is a great deal of slippage between espoused practices and complex student learning.

Kennedy (1998) found a similar relationship between espoused beliefs and teachers' responses to vignettes. That is, she found that average responses differed across these types of questions and

that the correlation between them was also low. One vignette, for example, had three parts: (a) questions about what was important for students to learn, (b) a question asking teachers to interpret a sample of student work, and (c) a question asking teachers how they would respond to this student if the student were in their own classroom. The first part asked for espoused ideas, while the second two parts were situated in a particular piece of student work. Kennedy found that teachers' espoused ideas tended to be more progressive and to include more complex forms of student learning, while their responses to the situated vignettes tended more toward providing students with rules and prescriptions. Kennedy noticed that, as the interviewers' questions moved closer and closer to the specific action of teaching (from an espoused idea to an interpretation of a student text to a response to the student text), teachers' responses became more and more prescriptive. Moreover, there was tremendous slippage from one level to the next. Findings such as Kennedy's and Oliver's suggest that third-level approximations are best construed as tapping teachers' espoused principles of teaching and that they may have only a modest relationship to teachers' likely practices and an even smaller relationship to complex student learning.

With respect to espoused ideas and achievement test scores, Peterson, Fennema, Carpenter, and Loeffel (1989) presented some data from a study design similar to Oliver's. They devised a set of principles from research on student learning in mathematics and translated these principles into questionnaire items. Teachers indicated the extent to which they agreed with each statement. Peterson et al. found that there was almost no relationship between teachers' espoused beliefs and the "number facts" portion of students' achievement test scores; there was, however, a modest (.32) relationship between teachers' espoused beliefs and the problem-solving portion of their students' achievement test scores. Cohen and Hill (1998), too, examined teachers' espoused principles of practice and student achievement test scores. They found a correlation between espoused principles and achievement scores of .29, very similar to the relationships to first-level approximations that were found by Oliver and by Peterson et al.

Fourth-Level Approximations: Testimony About Effects of Policies or Programs

Although third-level approximations seem distant from concrete teaching practices and even more

distant from complex student learning, there is one form of evidence that seems even more removed from these outcomes. What distinguishes the fourth-level approximation from the third-level approximation is the content of teachers' testimony. In a third-level approximation, teachers testify about the beliefs or principles of practice to which they subscribe, whereas, in fourth-level approximations, they testify about whether a particular policy or program was helpful or whether it influenced them in some way. Often, what is measured is teachers' satisfaction or dissatisfaction with a policy or a program, along with associated claims that the policy has helped or hindered their work. Sometimes, in addition, such claims are accompanied by estimates of how well the respondent believes he or she actually understands the policy in question. Even these questions, though, do not measure actual understanding; rather, they measure a self-assessment of understanding.

An example of a study depending on fourth-level approximations is Heneman's (1998) study of teachers' responses to a performance award system. Heneman was evaluating an incentive program that rewarded entire schools for school-wide improvements in student achievement rather than singling out individual teachers. He wanted to know whether such a system increased teachers' motivation. Since achievement data were clearly available in the school system under investigation, Heneman could have examined the achievement data themselves to determine whether there were increases in student learning over time. Instead, he surveyed teachers and asked questions such as those shown in Figure 4.

Fourth-level approximations are even more susceptible to self-serving biases than are second- and third-level approximations. One can imagine a teacher whose practice changes significantly toward compliance with a new policy but who insists in testimony that the policy is a major hindrance and is impossible to follow. In fact, efforts to find relationships between employee satisfaction and employee productivity have not been conclusive in either business or education (Johnson, 1990). However, to the extent that these testimonials move from the general to the particular, they may yield specific examples of policy hindrances or facilitators that are plausible. Johnson (1990), for instance, reported testimonials to the effect that teachers in certain schools could not obtain specific materials they needed to teach and that they had so much

[Sample from Section I.]

(Respond on a five-point agree-disagree scale:)

1. I have a clear understanding of what my school's benchmark goals are.
 7. It's hard to take the benchmark goals seriously.
 8. I am strongly committed to pursuing my school's benchmark goals.
- etc.

[Sample from Section II.]

In order to achieve your school's benchmark goals, to what extent have you:

1. Spent more hours on teaching and teacher-related tasks?
 2. Changed the content of what you taught?
 3. Changed the way you yourself taught (e.g., used different teaching methods)?
- etc.

12. How likely is it that you will exert the same level of effort to achieve your school's benchmark goal next year as you did this year?

- Very likely
 - likely
 - about as likely as unlikely
 - unlikely
 - very unlikely
- etc.

FIGURE 4. Sample of survey questions from the Heneman study of teacher incentives fourth-level approximation relying on testimony of policy influence).

water leakage near electrical outlets that they were afraid to plug in their computers. Such testimonials are more persuasive because they provide details about how the policy context directly influences classroom practices. Many studies relying on fourth-level approximations, though, do not elicit specific examples to support teachers' claims.

On the other hand, many of these researchers may not be interested in the effects of policies on teaching practices. Some have argued that studies using these testimonials do not intend to draw inferences about the processes of teaching and learning. Instead, they aim to draw inferences about employee job commitment, which in turn may affect such outcomes as daily attendance, attrition, or compliance with policies (Conley, 1991; Firestone, 1996). If this is the motivation for obtaining testimonies about satisfaction with policies, it may not be appropriate to list these testimonies

as fourth-level approximations to outcomes involving teaching and learning, particularly complex learning. Perhaps, instead, they are first- or second-level approximations of an elusive indicator of employee commitment. Still, much of this literature is presented in the same vicinity as literature concerned with improving the quality of teaching and learning, and often it is unclear whether authors are envisioning a path of influence that moves from policies to satisfaction, and then to teaching and learning, or whether, instead, they are envisioning a path that leads from policies to satisfaction, and then to commitment, and, say, attrition. Without clearer specifications of presumed paths of influence, these studies are likely to be understood as aiming for improved teaching and learning.

Still another problem with fourth-level approximations is that they often require teachers to compare their current knowledge or practices with their

recollection of their own prior knowledge or practices and to indicate whether or how much change has occurred as a result of the program or policy at issue. There is evidence to suggest that teachers are not particularly adept at accurately estimating their level of ability (and especially prior ability levels). For example, Strang, Badt, and Kauffman (1987) measured teachers' skills both before and after a training program, but they also asked the teachers to estimate their own proficiency before and after the program. The researchers' independent assessment of teacher change showed proficiency levels moving from 52% to 87% on their performance scale. However, the teachers' assessments of their own change indicated movement from 81% to 85%. Teachers, therefore, may not be good judges of whether they have learned from a program or have changed in response to a policy.

Relationship to Closer Approximations

There are many reasons to suspect that teachers' testimonials about the state of their own knowledge or about the influence of particular programs or policies on their practice may not accurately reflect their actual teaching practices or what students are actually learning in their classrooms. More than closer approximations, these fourth-level approximations invite self-serving commentary. In addition, because the terms are not situated, they invite different definitions.

In fact, I suspect that the distance between fourth-level approximations and closer approximations could account for some of the differences in findings across research studies. For instance, when Porter, Smithson, and Osthoff (1994) studied the impact of state curriculum policies, they found that teachers, for the most part, readily complied with new policies. But they also noted that their findings differed from those of Rosenholtz (1987). Rosenholtz's teachers indicated in their testimony that, although they sometimes complied, they were very reluctant to do so and were unhappy about it. These differences in teachers' responses could reflect differences in the types of data sought by the researchers. Porter et al. relied on second-level approximations—teacher logs and interviews—to learn what teachers were actually teaching, whereas Rosenholtz relied on fourth-level approximations: testimonials about teachers' satisfaction with the policies.

Testimonials also suffer because teachers may use terms differently than researchers do, or they

may use different yardsticks to gauge similarities and differences. Stigler and Herbert (1997), for instance, indicated surprise at the number of teachers in their sample who testified that they were familiar with mathematics standards developed by the National Council of Teachers of Mathematics (1988) and who testified that they had altered their teaching practices in response to those standards. Stigler and Herbert's classroom observations suggested that teachers' practices were very different from those implied by the standards. Other researchers have made similar observations about the disparity between teachers' claims to have adapted or adopted new policies and observers' independent assessments of the teachers' practices (e.g., Cohen, 1990). Applebee (1991) has suggested that such discrepancies exist because teachers incorporate reform ideas into their ongoing practices in such a way that the essence of the original idea is lost or distorted.

The Significance of Approximations

My focus on approximations in this article should not be construed as meaning that other dimensions of data collection devices are less important. Schwarz (1999) for instance, has shown the variety of ways in which data collection instruments can "lead the witness," so to speak, subtly conveying the sorts of responses that are expected. And Anti1 et al's (1998) study of cooperative groups suggests that careful definitions of terms are important. Many factors contribute to the quality of data collection devices. But we have attended much less to the problem of approximation than we have to some of these other issues.

Levels of approximation are important for at least two reasons. First, even apart from their degree of approximation to the elusive indicator of complex student learning, these different approaches to data collection each present different packages of advantages and disadvantages associated with cost, ease of use, availability, feasibility, susceptibility to errors, and so forth. I have tried to summarize these in Table 1. Each approach represents a trade-off among these many considerations in addition to their level of approximation. No doubt, each also involves other advantages and disadvantages that I have not included here.

Second, to the extent that researchers ultimately aim to learn how policies influence the intellectual character of classroom events or how they influence the quality of student learning, they need to

TABLE 1
Strengths and Weaknesses of Different Approximations

Source of data	Illustrative studies	Advantages	Disadvantages
Indicators of complex student learning	Shavelson (1983)	Face validity Curriculum sensitivity	No consensus on instrumentation or procedures Reliability hard to establish Expensive
First-level approximations			
Standardized achievement tests	Process-product research Education productivity research	Inexpensive Readily available Ubiquitous	Does not offer face validity Narrow range of outcomes
Observations of classroom lessons	Firestone et al. (1998) Newmann et al. (1996)	High face validity Plausible path to student learning	No standardized instruments Issues of sampling not worked out Labor intensive Disagreements about definitions of events
Second-level approximations			
Daily logs	Content determinants (Porter, 1989; Schwille et al., 1983)	Situated in teachers' own classrooms Can be aggregated Enable group comparisons	No shared meaning for terms Susceptible to self-serving biases Possibly burdensome for respondents
Vignettes describing hypothetical situations	TELT study (Kennedy, 1998) Floden et al. (1987) Ma (1999)	Standardized situations Can be aggregated Enable group comparisons Enable inclusion of teacher candidates	Unknown relation to practice Susceptible to self-serving biases
Third-level approximations			
Espoused principles and practices	McLaughlin (1993) Bidwell et al. (1997)	Easy and inexpensive to obtain May have symbolic value	Espoused principles often unrelated to practice No shared meaning of terms Susceptible to self-serving biases
Fourth-level approximations			
Testimony about effects of policies or of professional development programs	Heneman (1998) Rosenholtz (1987)	Easy and inexpensive to obtain May be useful for advocacy or for organizational development	No evidence of validity No shared meaning of terms Susceptible to self-serving biases Susceptible to estimation errors

show that there is a relationship between the approximations they choose to document and these outcomes. One reason that distant approximations present a problem for education policy researchers is that we lack an agreed-upon model that stipulates a path of influence moving from, say, a third- or fourth-level approximation to a first-level approximation or, better still, to an indicator of complex student learning. Absent such a model, correlational data are often used to demonstrate that there is at least some relationship among different levels of approximation. Many of the studies I have reviewed here provide some sort of evidence for relationships among different levels of approximation. Although this review is not intended to be exhaustive, a summary of these findings serves to illustrate the problem of defining a plausible path of influence that moves all the way to complex student learning.

Table 2 summarizes some of the correlations I have referred to in this article. Following Shavelson (1983), one might stipulate that a “substantial” predictive validity for any of these approximations would entail a correlation coefficient in the .50s. With that as the criterion, let us then examine the relationships shown in Table 2.

The first point to note in Table 2 is that the only relationships meeting Shavelson’s definition of substantial predictive validity are those that associate observed classroom practice with complex student learning, thus suggesting that they are indeed better approximations than are standardized tests. The second point to note, though, is that standardized achievement tests also have relatively high correlations with indicators of complex learning, even though, presumably, they are measuring a much narrower range of student learning outcomes. While these two first-level approximations are the closest approximations to indicators of complex student learning, Table 2 shows that most other approximations are of unknown value with respect to the elusive, ideal indicator of complex student learning.

Table 2 also shows relationships among more distant approximations. For instance, it suggests that third-level approximations tend to correlate in the low .30s with first-level approximations, standardized test scores, and classroom observations. Porter, Kirst, Osthoff, Smithson, and Schneider (1993) obtained some higher correlations (between questionnaire data and log data), but these higher correlations involved questions about topics, and cor-

relations were lower when questions dealt with the type of intellectual work being done within a given topic. The very diversity of correlation coefficients, however, indicates that even within a level of approximation, there can be remarkably different predictive power depending on how the question was put, how teachers understand the terms, to what extent the response involves presentation as opposed to representation (Freeman, 1996), and so forth. The value of these third-level approximations has to be considered both in light of this variability in relationships to closer approximations and in light of the fact that classroom observations and standardized tests are themselves merely approximations of complex student learning. As a whole, the pattern of correlations suggests that, as researchers move to more distant approximations, they are gathering evidence on outcomes that may be only weakly related to the outcomes policymakers ultimately hope to influence.

Finally, Table 2 indicates that no one (of whom I am aware) has attempted to measure the relationship between fourth-level testimonials about policy impact and any closer levels of approximation. The significance of this void is not clear, since the analysis presented here is based on an assumption that education policy researchers generally want to estimate the chances for improved student learning and that, in particular, they are interested in more complex forms of student learning. Yet, it is possible that those who rely on fourth-level approximations are more interested in teacher compliance, endurance, or malleability than in student learning. Thus, the lack of knowledge about relationships between fourth-level approximations and, say, first- or second-level approximations could reflect the fact that researchers focusing on fourth-level approximations are not really aiming to approximate student learning but instead are interested in approximating something like employee commitment or endurance.

There is one further important point that needs to be made about these different levels of approximation: They are not related to the popular methodological distinctions we make between quantitative and qualitative research methods. In fact, I have included in this article both quantitative and qualitative examples of almost all levels of approximation. Since both types of methods can be used at virtually any level of approximation to the elusive indicator of complex student learning, I suspect that the distinction between these broad cat-

TABLE 2
Correlations Among Different Levels of Approximation

Level of approximation	Direct indicators of complex learning	Level 1 approximations		Level 2 approximations
		Standardized test scores	Classroom observations	Situated descriptions
<i>First level</i>				
Classroom observations	.59 ^a ; .77 or curvilinear, depending on students ^b			
Standardized achievement tests	.43 ^c			
<i>Second level</i>				
Teacher logs			Focused reports agree in the .70s with observations ^d	
Focused self-reports				
<i>Third level</i>				
Interviews		.29-.33 ^e	.30; average scores more progressive than observed practice ^f	.02-.93 when questions are about topics, - .01-.48 for questions about intellectual work ^g ; espoused principles less directive than vignette responses, relationships also low ^h
<i>Fourth level</i>				

Note: Studies in the direct indicators column used quite different indicators of complex student learning.

^a Newmann et al. (1996). (Number shown here is the square root of their R^2 . It is the only estimate shown derived from a multiple regression rather than from a simple correlation coefficient.)

^b Based on a hierarchical linear model of Saxe et al. (1999).

^c Shavelson et al. (1992).

^d Koziol and Burns (1986).

^e Cohen and Hill (1998), Peterson et al. (1989) Elliott (1998).

^f Oliver (1953).

^g Porter et al. (1993).

^h Kennedy (1998).

egories of methods is less important than is the level-of-approximation distinction, which has to do with how closely the documented outcomes approximate the ultimate policy outcomes of interest.

Conclusion

My aim in this article has not been to present a comprehensive review of policy research; rather, I have attempted to draw on this literature to raise questions about how different kinds of evidence might be interpreted in terms of the ultimate policy aim of improving complex student learning. The correlations among different levels of approximation that I have cited here are suggestive of the magnitude of the approximation problem. These correlations suggest that, absent an agreed-upon indicator of complex student learning, the best first-level approximation available to researchers would be a classroom observation focusing on the nature of intellectual work students do in class. However, observations are extremely labor intensive and, hence, may not be feasible for many researchers with tight budgets. Thus, the most important consideration for many researchers may be how to develop questions that can be used in interviews or questionnaires.

Although the literature is scanty, it does suggest that researchers would be wise to distinguish between situated and nonsituated teacher testimonials, since the nonsituated statements do not seem to be very highly related to closer approximations. For those policy researchers who are not able to use direct observations, therefore, it might prove fruitful to invest time in developing and field testing questions that are closely situated so that the distance between these questions and closer approximations is reduced as much as possible. More than anything, this brief review suggests the need for a clearer understanding of how these different levels of approximation are related to one another, not only correlationally but causally.

References

- Antil, L. R., Jenkins, J. R., Wayne, S. K., & Vasdasy, P. F. (1998). Cooperative learning: Prevalence, conceptualizations, and the relation between research and practice. *American Educational Research Journal, 35*, 419-454.
- Applebee, A. N. (1991). Informal reasoning and writing instruction. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 225-246). Hillsdale, NJ: Erlbaum.
- Argyris, C., & Schon, D. A. (1996). *Organizational learning II*. New York: Addison-Wesley.
- Bidwell, C. E., Frank, K. A., & Quiroz, P. A. (1997). Teacher types, workplace controls, and the organization of schools. *Sociology of Education, 70*, 285-307.
- Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 328-375). New York: Macmillan.
- Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis, 12*, 311-329.
- Cohen, D. K., & Hill, H. C. (1998). *Instructional policy and classroom performance: The mathematics reform in California*. Philadelphia: Consortium for Policy Research in Education.
- Conley, S. (1991). Review of research on teacher participation in school decision making. In G. Grant (Ed.), *Review of research in education* (Vol. 17, pp. 225-266). Washington, DC: American Educational Research Association.
- Firestone, W. A. (1996). Images of teaching and proposals for reform: A comparison of ideas from cognitive and organizational research. *Educational Administration Quarterly, 32*, 209-235.
- Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis, 20*, 95-113.
- Floden, R. E., Porter, A. C., Schmidt, W. H., Freeman, D. J., & Schwille, J. R. (1987). Responses to curriculum pressures: A policy capturing study of teacher decisions about content. *Journal of Educational Psychology, 73*, 129-141.
- Freeman, D. (1996). "To take them at their word": Language data in the study of teachers' knowledge. *Harvard Educational Review, 66*, 732-761.
- Greenwald, R., Hedges, L., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66*, 361-396.
- Heneman, H. G. III. (1998). Assessment of the motivational reactions of teachers to a school-based performance award program. *Journal of Personnel Evaluation in Education, 12*, 43-59.
- Irving, R., & Elton, M. C. J. (1986). The use of diaries to measure discretionary behavior: Hypotheses and results. *Evaluation Review, 10*, 95-113.
- Johnson, S. M. (1990). *Teachers at work: Achieving success in our schools*. Boston: Basic Books.
- Kennedy, M. M. (1998). *Learning to teach writing: Does teacher education make a difference?* New York: Teachers College Press.
- Koziol, S. M., & Burns, P. (1986). Teachers' accuracy in self-reporting about instructional practices using a focused self-report inventory. *Journal of Educational Research, 79*, 205-209.

- Ma, L. (1999). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States*. Mahwah, NJ: Lawrence Erlbaum.
- McLaughlin, M. (1993). What matters most in teachers' workplace context? In J. W. Little & M. W. McLaughlin (Eds.), *Teachers' work* (pp. 79-103). New York: Teachers College Press.
- National Council of Teachers of Mathematics. (1988). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.
- Newmann, F. M., Marks, H. M., & Gamoran, A. (1996). Authentic pedagogy and student performance. *American Journal of Education*, 194, 280-312.
- Oliver, W. A. (1953). Teachers' educational beliefs versus their classroom practices. *Journal of Educational Research*, 48, 47-55.
- Peterson, P. L., Fennema, E., Carpenter, T. P., & Loef, M. (1989). Teachers' pedagogical content beliefs in mathematics. *Cognition and Instruction*, 6, 1-40.
- Porter, A. C. (1989). A curriculum out of balance: The case of elementary school mathematics. *Educational Researcher* 18(5), 9-15.
- Porter, A. C., Kirst, M. W., Osthoff, E. J., Smithson, J. L., & Schneider, S. A. (1993). *Reform up close: An analysis of high school mathematics and science classrooms*. Madison: Wisconsin Center for Education Research.
- Porter, A. C., Smithson, J., & Osthoff, E. (1994). Standard setting as a strategy for upgrading high school mathematics and science. In R. F. Elmore & S. H. Fuhrman (Eds.), *The governance of curriculum* (pp. 138-166). Alexandria, VA: Association for Supervision and Curriculum Development.
- Rosenholtz, S. J. (1987). Education reform strategies: Will they increase teacher commitment? *American Journal of Education*, 95, 534-562.
- Rossi, P. (1979). Vignette analysis: Uncovering the normative structure of complex judgements. In R. K. Merton, J. S. Coleman, & P. H. Rossi (Eds.), *Qualitative and quantitative social research: Papers in honor of Paul Lazarsfeld* (pp. 175-188). New York: Macmillan.
- Saxe, G., Gearhart, M., & Seltzer, M. (1999). Relations between classroom practices and student learning in the domain of fractions. *Cognition and Instruction*, 17, 1-24.
- Schwarz, N. (1999). Self reports: How the questions shape the answers. *American Psychologist*, 54, 93-105.
- Schwille, J., Porter, A., Belli, G., Floden, R., Freeman, D., Knappen, L., Kuhs, T., & Schmidt, W. (1983). Teachers as policy brokers in the content of elementary school mathematics. In L. S. Shulman & G. Sykes (Eds.), *Handbook of teaching and policy* (pp. 370-391). New York: Longman.
- Shavelson, R. J. (1983). Review of research on teachers' pedagogical judgements, plans, and decisions. *Elementary School Journal*, 83, 392-413.
- Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher* 21(4), 22-27.
- Shavelson, R. J., Web, N. M., & Burstein, L. (1986). Measurement of teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 50-91). New York: Macmillan.
- Stigler, J. W., & Herbert, J. (1997). Understanding and improving classroom mathematics instruction. *Phi Delta Kappan*, 79, 14-21.
- Strang, H. R., Badt, K. S., & Kauffman, J. M. (1987). Microcomputer-based simulations for training fundamental teaching skills. *Journal of Teacher Education*, 38, 20-26.

Author

MARY M. KENNEDY is a professor at Michigan State University, Department of Teacher Education, 116F Erickson Hall, East Lansing, MI 48824. She specializes in teacher knowledge and teacher learning.

Manuscript received January 12, 1999

Accepted June 14, 1999