A university professor was on her hands and knees in the darkest part of a tunnel when a sewer inspector stumbled over her.

“What are you doing here?” he asked.
“I’m looking for the truth, or at least useful knowledge,” was the reply. “I’m pretty sure someone must have left it here.”
“But it’s so dark, muddy, and toxic here that I fear you will never find anything of value. Why don’t you go look where there is a bit more light?”
“Because the muddier it is, the better the chance that I will find something truly interesting. It’s so unlikely that anything of value will be hidden in plain sight.”

What kind of research on teaching is of most worth? To what extent should researchers in this field be conducting highly functional investigations that attempt to identify the key elements of accomplished teaching or the most important components of teacher preparation programs or experiences? Should we be asking whether teacher education programs significantly improve the likelihood that someone will teach effectively? Should we instead be conducting inquiries that explore the rich complexities of teaching, learning, schooling, and development and the contexts that support them? What genres of research are worth undertaking?

The tacit dialogue between the present articles by Wilson, Floden, and Ferrini-Mundy (2002 [this issue]) and by Florio-Ruane (2002 [this issue]) is nostalgically familiar. We designed the Institute for Research on Teaching (IRT) in 1975 on the basis of our critique of the then-prevailing prototype of process-product research on teaching. We considered process-product research on teaching behavioristic, simplistic, and unduly dependent on standardized achievement tests as indicators of product. Indeed, the leaders of process-product research, such as Nate Gage (1978) and Barak Rosenshine (Rosenshine & Stevens, 1986), complained that their critics were unnecessarily “complexifying” the phenomenon of teaching, whereas the hallmark of scientific progress was increased simplification, not complication. Moreover, if research on teaching were to have the desired impact on policy makers, it needed to be both simple and clearly connected to easily understood indicators of student achievement. Finally, there was a moral message in the process-product tradition. Our bottom-line obligation as teachers was to the students and their learning. To study teaching without reference to students was unethical self-indulgence.

These two articles stimulated me to reflect on my history of work as an active scholar on teaching and teacher education. I thought about the nearly four decades of research in which I had been actively involved. And I began to wonder how, if at all, it added up.

I concluded that we may be asking the wrong questions and focusing on the wrong units of analysis. That is, individual studies rarely can be adjudged as valuable or trivial per se. Instead, we need to think about extended programs of scholarship, in which a variety of types of research are pursued, to maximize the value to be gained from studies of teaching. I want to tell a story of more than 30 years of research, of a series of research programs that cumulated into a meaningful knowledge base, an enduring policy initiative, and the spinning off of a number of significant lines of research.

I begin with my work on medical problem solving in the 1960s and 1970s, followed by the research on teaching as information processing that characterized the IRT programs. A set of
studies on the development of teacher knowledge, with special reference to pedagogical content knowledge, followed that work, which transitioned into the Teacher Assessment Project conducted on behalf of the then-infant National Board for Professional Teaching Standards. The board’s own validation studies of board certification were conducted in the year 2000, and a new program of Carnegie Foundation studies of teacher education is currently underway. This sequence will help to illustrate my conception of the value of general programs of research that alternate freely among genres and constitute, in Darwinian retrospect, an adaptively grand strategy.

MEDICAL PROBLEM SOLVING

In 1968, Arthur Elstein and I began a long series of studies of medical problem solving at Michigan State University. I had earlier been studying problem solving among preservice teachers. Michigan State University had just begun an innovative new medical school, with a commitment to a problem-solving-centered curriculum. Our research was embedded within the work of inventing and implementing this new curriculum. We sought to gather empirical evidence of what outstanding medical diagnosticians actually did, to provide a more substantial basis for defining both processes and outcomes in this new curriculum. We also felt that if we could develop a better understanding of how medical diagnoses were accomplished, we would make a fundamental contribution to the understanding of all complex human cognition. In that way, I hoped, I could ultimately return to understanding teacher problem solving.

Among our most interesting (and most surprising) findings was that there was no evidence of a generic diagnostic competence. Instead, medical diagnosis was domain specific (Elstein, Shulman, & Sprafka, 1978). We developed methods to study professional thinking through a combination of thinking aloud procedures in real time plus the use of videotape-stimulated recall to reconstruct thought processes after the fact. We also developed content-analytic techniques that would become invaluable in our later studies of teacher thinking and the wisdom of practice (Shulman & Elstein, 1975). These studies became the basis for a new theory of medical problem solving. They also became a starting point for our subsequent studies of teaching.

THE IRT: “IT’S THE THOUGHT THAT COUNTS”

Why were physicians treated as autonomous problem solvers whose thinking needed to be understood and supported, whereas teachers—in the early 1970s—were conceptualized as mere emitters of behavior? Wasn’t it likely that we would make far more progress in understanding and improving teaching through studying teacher thinking, planning, decision making, and judgment? When the opportunity arose to compete for the federally funded IRT in 1975, we adapted the models and methods developed for the study of physicians and proposed to apply them to the study of teaching. Moreover, we engaged anthropologists such as Susan Florio-Ruane and Fred Erickson to join us at the IRT. Ethnographers were also concerned with the motives and implicit reasoning that explained teacher and student behavior.

We also decided to include teachers as active collaborators in every research program at the IRT, buying out half the contract time of eight local teachers so they could spend substantial energy as in-house coinvestigators. We had never dreamt of conducting research on medical thinking without the active collaboration of physicians, so how could we do otherwise with teaching? We did not fully realize at the time how revolutionary that policy would become.

The IRT research program, which continued for nearly a decade, helped to redirect the paradigms of research on teaching by shifting attention from behavior to thought; from observable
performance to strategy and understanding; and from simple models of stimulus and response to more complex and subtle models involving context, content, and cognition. But we also managed to lose sight of a key principle of product-process research: the importance of linking the distinctive features of teaching to the quality of student learning. In rejecting the validity of standardized achievement tests as adequate representations of student understanding, we also disconnected our complex accounts of teaching from commonly accepted indicators of student learning. Baby was discarded along with bath water. By disdaining the value of measures that were readily visible where the light was strong, we lost sight of the strategic and moral need to see the consequences of teaching in some straightforward manner. We became so enamored of teachers’ cognitive processes that we ignored student learning products almost entirely, an error that would prove costly to researchers on teaching in dialogue with the policy community 20 years later.

**TEACHER KNOWLEDGE**

In the early 1980s, having moved from Michigan State to Stanford, my students and I began asking not, How do teachers think and make decisions? but, What do teachers know and how do they use what they know? More specifically, we asked how teachers, who already know and understand their subjects in particular ways, learn to transform their knowledge into representations that make sense to their students. This new line of work, growing out of the earlier studies in medicine and teaching, both substantively and methodologically, now placed much greater emphasis on teachers’ content knowledge and their content knowledge for teaching, which we dubbed “pedagogical content knowledge.” Connecting back to the earlier medical studies, we also posited that teachers’ knowledge for teaching was content specific, that what teachers needed to know to teach math to 2nd graders, for example, was quite different from the knowledge needed to teach history to 10th graders (Shulman, 1986, 1987).

A program of dissertation research was connected to the core program. For example, Pam Grossman (1989) studied how English teachers who were pedagogically prepared in particular ways approached and accomplished the teaching of Shakespeare differently from Shakespeare-sophisticated teachers who were not so prepared. Maher Hashweh (1985) documented how teachers specializing in a particular science (biology or physics) approached the explanation of new concepts to middle school students in their familiar and their unfamiliar content areas. Bill Carlsen (1988) demonstrated how teachers of biology taught topics that they understood well quite differently from topics they understood more marginally. Sigrun Gudmundsdottir, Suzanne Wilson, Rick Marks, Sam Wineburg, and others conducted studies of teaching and understanding history, mathematics, literature, and other subjects. In general, while studying the learning of new teachers as they progressed from preservice to the first years of teaching, the teacher knowledge project helped to locate the concepts of content and pedagogical content knowledge as central features of the working model of teacher expertise. Indeed, the content specificity of teaching was a fundamental principle. However, the teacher knowledge projects did not seek to document connections between teacher knowledge and student achievement.

**TEACHER ASSESSMENT PROJECT**

When the opportunity arose to define and design a national board for accomplished teachers, our group at Stanford conducted the research and development needed to create prototypes of the board certification instruments. From 1985 to 1990, our research group thus shifted from the exploratory, theory-driven, basic research characteristic of the IRT and teacher knowledge research to more product-
oriented investigations: creation of a national board assessment program. This work built directly on the earlier teacher knowledge research. It was predicated on a conception of teaching consistent with our theoretical model of pedagogical reasoning and action. It represented teaching as resting on deep content knowledge, on pedagogical content knowledge as the basis for transforming teacher understanding into pedagogical representation, on the ability to reflect on and learn from one’s own teaching experiences, and on the assumption of subject-specific pedagogy.

Thus, each assessment began with wisdom-of-practice studies of accomplished teachers, using their thinking and action as a basis for establishing standards and scoring rules. From the very beginning, the development work on the board assessment presumed that different assessments would be necessary for the teachers of different subjects to different ages. The assessment would strongly emphasize the work of teachers as designers of, assessors of, and reflectors on teaching in addition to their observed performances in the classroom. The assessment would both tap into teacher knowledge and require that the knowledge be demonstrated, whenever possible, in the context of the classroom in which they were actually teaching. Thus, the theoretical constructs that had developed within the basic research—content, cognition, and context—appeared at the heart of the policy-relevant development of board certification. The applied work would have been impossible without the decade of more basic work that preceded it. Moreover, even in the heat of the assessment development, important fundamental studies were conducted as well (e.g., Wilson & Wineburg, 1993).

There were also significant new methodological advances. None was more important than the work we conducted on the design, use, and evaluation of classroom teaching portfolios. This work, pursued in the interests of test development, would ultimately have significant influence on teacher education, its pedagogy, and its assessment, more broadly.

But if our national board work reflected the theoretical importance of the three Cs of content, cognition, and context, it continued to ignore a fourth C, consequences for students. Although the contextualized, content-specific portfolio entries were expected to include evidence of student learning, systematic connections between the features of board certification and student achievement remained undemonstrated.

**THE NATIONAL BOARD VALIDATION STUDY: LLOYD BOND’S RESEARCH**

The national board has become one of the most durable policy instruments for the reform of American education through improving teacher quality. Between 1990 and 2000, the national board grew from a vision and a set of prototypes into an institution. By fall 2001, more than 16,000 teachers had been certified from every state, more than 6,000 in 2001 alone. A target of 100,000 board-certified teachers by 2006 had been announced. The assessment process, a combination of a context-based teaching portfolio and a computer-based assessment center, was being administered in more than 30 specific domains. And finally, under the leadership of Lloyd Bond, the first comprehensive validation study was conducted (Bond, Jaeger, Smith, & Hattie, 2000). Bond sought to connect process with product. In Bond et al.’s (2000) study, teachers who had scored in four different quartiles on the board assessments were observed in their classrooms by raters who had no knowledge of their board performance. Samples of their students’ writing on specified units were independently evaluated. The study demonstrated that the higher the teachers scored on the board assessment, the closer their actual classroom teaching corresponded to the board’s
standards of accomplished teaching, and the more highly their students achieved in their writing assignments.

Thus, a model of teaching that developed out of a critique of process-product research and was based on wisdom-of-practice studies of both teaching and medicine was elaborated in further studies of teacher knowledge development among those learning to teach and became the basis for the design and implementation of a radically new approach to evaluating the quality of teachers. This assessment was then validated against observed classroom teaching performance and measures of student learning.

Validity questions must continue to be raised about both competing forms of teacher assessment and alternate approaches to teacher preparation and selection. But this account of a flow of research programs illustrates clearly, I hope, the utter interdependence among these genres of research when looking at the big picture over time. Seeking more light and seeking practical solutions are not antithetical when viewed programmatically.

A TOOLKIT FOR THE PREPARATION OF TEACHERS

My colleagues at the Carnegie Foundation and I are once again embarked on a study whose purpose is both to understand and to improve the quality of teacher education, however undertaken. We have decided that it makes little sense to ask the recurring question, Does teacher education make a difference? or, Is preparation in a formal teacher preparation program superior to no teacher education (or a diminished form thereof) at all? Neither teacher education nor any of its alternatives defines a treatment, an independent variable, or any other intervention with any precision. Indeed, all forms of teacher education, wherever they occur, may be considered alternative programs. Unlike preparation for law or medicine, there exists no standard model for the preparation of teachers, even within accredited university programs.

Instead, we have concluded that the understanding of teacher education, much less its serious improvement, will be unlikely to occur without development and field testing of a set of instruments for documenting and measuring the various important dimensions of both teacher learning and development and of the opportunities to learn that are critical for future teachers in any track. Our current conclusion is that the field is in serious need of low-stakes, high-yield instrumentation to monitor the vital signs of teacher development in ways that can guide teacher educators, professional developers, and ultimately teachers themselves. Some of these instruments may look like tests, others like portfolio entries or stations in assessment centers. But we are embarked, in collaboration with a small number of teacher education programs, on a program to develop, field test, and experiment with a teacher educator’s “toolkit.”

These instruments must permit the eventual study of connections between capacities that teachers can intentionally and effortfully learn to develop—wherever they occur—and the impacts of their efforts on the learning and development of students. But these connections need not, indeed ought not, characterize each study conducted. These are characteristics of a full program of research, not each individual study.

CONCLUSIONS

What then do I conclude from this reminiscence about several decades of research on teaching and teacher education?

1. We need no more studies that compare the presence of teacher education with its absence. These are not worthwhile studies. If policy makers continue to insist that they want such studies, ignore them. Studies not worth conducting are not improved through meta-analysis. We need instead studies that compare meaningful, credible, alternative designs. Remember Cronbach’s admonition that if you want to compare a camel with a horse, compare the best camel you can find with the best horse you can find. Do not get two camels and cut the hump off one of them.

2. Large-scale generalizations about the efficacy of teacher education are unlikely to result from empirical research. Nevertheless, this does not obviate the importance of good instrumentation, careful measurement, scrupulously faithful ethnographic accounts, and carefully reasoned inferences and arguments. Local knowledge needs precision as much as
classic scientific inquiries. As Geertz (1973) observed, citing Robert Solow, just because complete antisepsis is impossible does not mean that we should perform surgery in a sewer.

3. Publication in peer-reviewed journals is no longer a meaningful benchmark for the value of a study. Much important work appears in edited volumes and in special reports. Increasingly, work will appear on Web sites. Many important journals (e.g., *Harvard Education Review, Kappan*) have never been peer reviewed. We need to develop new mechanisms for quality assurance in research on teaching.

4. The appropriate unit of analysis is not the individual study, except on those rare occasions when the study is of such magnitude that it represents a program of inquiry, not a single study in traditional terms.

5. We need programs of research that combine some of the following features: wisdom of practice, visions of the possible, instrument development and the “instrumenting” of programs or interventions to identify connections between opportunities to learn and enduring changes in teacher capacities, and establishing connections to student learning and development.

6. Whenever possible, studies of teacher learning should be embedded within ongoing programs of teacher education in the form of “scholarship-of-teaching-and-learning” programs that combine instrumentation, opportunities to learn, etc. in the form of quasi-design experiments (Hutchings & Shulman, 1999).

If teacher education research is to make a significant difference, we must make new scholarly commitments. Long-term programs of scholarship, regularly embedded in ongoing programs of teacher preparation, must become a norm, with ties to student learning established wherever and however possible. If we wish to live in a society enlightened about teaching and teacher education, we will need to look for those truths that have real consequences for teachers, students, and society and to find ways to shed the light of scholarship on many dark corners.

REFERENCES


Lee S. Shulman is the president of the Carnegie Foundation for the Advancement of Teaching and the Charles E. Ducommun Professor of Education Emeritus at Stanford University. A past president of both the American Education Research Association and the National Academy of Education, his research has focused on professional education in medicine and teaching, the logic of educational research, the cognitive psychology of instruction, and the scholarship of teaching in higher education.