FIVE

THE EVALUATION OF THE STANFORD TEACHER EDUCATION PROGRAM (STEP)

An Interview With David Fetterman

Introduction: David Fetterman was a member of the faculty of the School of Education and Director of the MA Policy Analysis and Evaluation Program at Stanford University at the time he conducted this evaluation. He is currently Director of Evaluation in the Division of Evaluation in the School of Medicine at Stanford University. He is the past president of the American Evaluation Association. Fetterman has received the highest honors from the association, including the Lasersfeld Award for evaluation theory and the Myrdal Award for evaluation practice. Fetterman has been a major contributor to ethnographic evaluation and is the founder of empowerment evaluation. He has published 10 books and more than 100 articles, chapters, and reports, including contributions to various encyclopedias. His most recent books include *Empowerment Evaluation Principles in Practice* and *Ethnography: Step by Step.* He is President of Fetterman & Associates, an international consulting firm, conducting work in Australia, Brazil, Finland, Japan, Mexico, Nepal, New Zealand, Spain, the United Kingdom, and the United States.

This interview concerns Fetterman's complex, three-year evaluation of the Stanford Teacher Education Program (STEP). In this evaluation, Fetterman chose to use an approach other than his well-known empowerment approach. He describes the reasons for his choice, which provides guidance as to the conditions necessary to use an empowerment approach. As a member of the Stanford University education faculty, though not a member of STEP, Fetterman served in a partially internal evaluator role and discusses some of the problems he encountered in that role. He describes the methods he used, including intensive immersion, surveys, interviews, reviews of literature, and discussions with experts in other teacher education programs, to judge the quality of delivery of STEP and some of the conclusions he reached.

Summary of the STEP Evaluation

David Fetterman

The president of Stanford University, Gerhard Casper, requested an evaluation of the Stanford Teacher Education Program (STEP). The first phase of the evaluation was formative, designed to provide information that might be used to refine and improve the program. It concluded at the end of the 1997–1998 academic year. Findings and recommendations from this phase of the evaluation were reported in various forms, including a formal summer school evaluation report (Fetterman, Dunlap, Greenfield, & Yoo, 1997), more than 30 memoranda, and various informal exchanges and discussions.

The second stage of this evaluation was summative in nature, providing an overall assessment of the program (Fetterman, Connors, Dunlap, Brower, Matos, & Paik, 1999). The final report highlights program evaluation findings and recommendations, focusing on the following topics and issues: unity of purpose or mission, curriculum, research, alumni contact, professional development schools/university school partnerships, faculty involvement, excellence in teaching, and length of the program. Specific program components also were highlighted in the year-long program evaluation report, including admissions, placement, supervision, and portfolios. (See the STEP Web site for copies of all evaluation reports: www.stanford.edu/davidf/step.html.)

The Methodology

The evaluation relied on traditional educational evaluation steps and techniques, including a needs assessment; a plan of action; data collection (interviews, observations, and surveys); data analysis; and reporting findings and recommendations. Data collection involved a review of curricular, accreditation, and financial records, as well as interviews with faculty and students, and observations of classroom activity. Informal interviews were conducted with every student in the program. Focus groups were conducted with students each

quarter and with alumni from the classes of '95, '96, and '97. More than 20 faculty interviews were conducted. Survey response rates were typically high (90%–100%) for master teachers, current STEP students, and alumni. Data collection also relied on the use of a variety of technological tools, including digital photography of classroom activity, Web surveys, and evaluation team videoconferencing on the Internet. Data analysis was facilitated by weekly evaluation team meetings and frequent database sorts. Formal and informal reports were provided in the spirit of formative evaluation. Responses to preliminary evaluation findings and recommendations were used as additional data concerning program operations. (A detailed description of the methodology is presented in Fetterman, Connors, Dunlap, Brower, & Matos, 1998.)

Brief Description of STEP

STEP is a 12-month teacher education program in the Stanford University School of Education, offering both a master's degree and a secondary school teaching credential. Subject area specializations include English, languages, mathematics, sciences, and social studies. The program also offers a Cross-Cultural, Language, and Academic Development (CLAD) emphasis for students who plan to teach second-language learners. The 1997–1998 class enrollment was 58 students. Tuition and board were approximately $30,000.

The program introduces students to teaching experiences under the guidance of a master teacher during the summer quarter. Students enter the academic year with a nine-month teaching placement, which begins in the fall quarter under the supervision of a cooperating teacher and field supervisor. Students also are required to take the School of Education master's degree and state-required course work throughout the year.

The program administration includes a faculty sponsor, director, placement coordinator, student services coordinator, lead supervisor, field supervisors, and a program assistant. In addition, the program has a summer school coordinator/liaison and part-time undergraduate and doctoral students.

Findings, Recommendations, and Impact

The most significant finding was that the STEP program had some of the ingredients to be a first-rate teacher education program, ranging from a world-renowned faculty to exceptional students. At the time of the evaluation, the

program and faculty had a unique opportunity to raise the standard of excellence in the program and the field.

The evaluation identified some noteworthy qualities of STEP. These included high-caliber faculty and students, supportive and critical supervision, the year-long student teaching experience, a culminating portfolio conference, and strong support from alumni. Nevertheless, problem areas were identified. Key among these was the lack of a unifying purpose to shape the program. Related to the absence of a clear vision for the program was the fact that faculty designed their courses in isolation from each other and the central activities of STEP, leading to a fragmented curriculum and a lack of connection between educational theory and practice. Instructional quality was occasionally a problem, particularly as students expect to have faculty they can view as models for exemplary teaching. Students also received no systematic research training to help them develop an inquiry-based approach to teaching. Finally, the program may need to be lengthened to accomplish all that is desired.

Final recommendations included developing a mission statement focusing on reflective practice; instituting faculty meetings and retreats to design, revise, and coordinate curriculum and course content; reducing fragmentation in the curriculum and developing a rationale for course sequencing, including more content on classroom practice to balance educational theory; developing a research training program; forging school-university partnerships; and adopting a commitment to excellence in teaching. The findings and recommendations made in this evaluation went beyond tinkering at the fringes of the program. Many recommendations represented significant and fundamental changes in the program.

The use of the evaluation was gratifying. More than 90% of the recommendations were adopted, ranging from small-scale curricular adaptations to large-scale programmatic redefinitions. The success of this evaluation helped launch the development of an undergraduate teacher education program as well. In addition to the impact the evaluation had at Stanford, the teacher education evaluation set a standard for teacher education programs nationally and internationally. The report was used and referenced almost immediately and served as the catalyst for changes in programs throughout the world.

Dialogue With David Fetterman

Jody Fitzpatrick

Fitzpatrick: David, you're known for your development of the empowerment evaluation approach, yet you chose not to use this model for the evaluation of STEP. Tell us a little bit about *why* you took a different approach here.

Fetterman: Well, there's a rational basis for my decision, but it also was informed by personal judgment and experience. The rational part of my decision was very simple: The president of Stanford requested the evaluation. His request was more like the traditional accountability focus of evaluation. There are multiple purposes for evaluation: development, accountability, knowledge. Empowerment falls more into the development purpose rather than into traditional accountability. If I truly believe that, that means I must abide by those distinctions and use the traditional approach when it is more appropriate. A lot of people think I do only empowerment, but I've done traditional evaluation for 20 years. These different approaches inform each other. The choice depends on the situation you're in. I think most evaluators believe that. You can't just be a purist or dogmatic about your approach.

Then, there's the trickier level beyond all that—the personal judgment part. For empowerment evaluation to work, I have to have a sense that there is both group cohesion and trust. Often, one has to make a guess about these traits at a very early phase in the study before you know as much as you'd like to about the context. But, at the early stage of this study, I didn't sense enough group cohesion and trust to proceed with an empowerment approach.

Now let me say, there were parts of the study where I used the empowerment approach, though 99% of it was traditional. I used a matrix tool and interviewed a lot of students to get their ideas about what was the most important part of the program to help us in planning the evaluation. That's not empowerment, but it is applying some of the concepts and techniques.

But the main reasons for not using empowerment were the chain of command—the president asked for the study—and my judgment of the nature of the place—whether the place has enough trust to do empowerment work.

There is a rational process that plays a clear, dominant part in one's choices, but then there's an important intuitive part. We have to admit it. As we get a little more seasoned in doing evaluations, our choices don't always emerge from a logical flow. Personal judgment is involved as well. The important thing is to make these judgments explicit and to link them to the rational. You triangulate to see if there's enough substance to the intuitive. You're just as critical about the intuitive as the rational.

Fitzpatrick: You describe the model you used as a traditional educational evaluation. Tell us a bit about the approach. Would you consider it theory based, decision oriented, goal-free, . . . ?

Fetterman: We began by trying to understand what the model for STEP was—what insiders said it was *supposed to do.* Then, we wanted to look at what kind of compliance there was with the model, but from a consumer perspective. That is, we came up with the basic program theory and then looked at it to see if the action linked up with the plan. It was a mature program and we wanted to look at what it was doing—how it was operating. If it were a new program, we would have used different criteria, but it had been around a while. We also were decision oriented and objectives oriented.

The directive from the president and provost was absolutely clear: This was a Stanford program. They wanted to know if it met that criterion. Was it up to Stanford standards? The president had been the provost at the University of Chicago when they eliminated the School of Education, so the stakes were perceived as extremely high. Many egos were at stake. That's why it was such an intense evaluation.

Fitzpatrick: It must have been an interesting process for you to use a different approach. I think we always learn from trying something new. What did you learn about evaluation from conducting this study?

Fetterman: The approach wasn't new for me because I've always done traditional evaluations along with participatory and ethnographic, but you always learn something new when you do any evaluation. I learned some things that were kind of scary and interesting methodologically. The complexity of the context was astounding. Even when you've known the organization for years, you learn new things about the context in an evaluation, sometimes things that you don't necessarily want to know! I was interviewing one colleague when suddenly, in the middle of the interview, the guy was almost in tears about being pushed away from teaching in the program. He was so hurt and shaken by that, and it had happened 20 years ago! He had been carrying

that around all that time. Many faculty members had been hurt by their association with this program. This was not an isolated incident. These feelings explained the landmines we encountered later. It helped explain why some of the responses to the evaluation were so strong.

One of the interesting methodological surprises we encountered was almost a Lake Wobegon effect on students' course evaluations. When you looked at just the survey results, all the faculty got a 3 or above on a 5-point scale in which 1 is *poor*, 3 is *satisfactory*, and 5 is *excellent*, but when you actually observed the classes and interviewed students, the variability in the assessments of teaching quality went from 0 to absolutely stellar. The surveys didn't capture anything below the satisfactory level and had a ceiling effect on the top performers. That was scary because most evaluations rely so much on surveys. If you don't supplement these, you can have a really false perception of the quality of teaching in the program. When you interviewed and observed classes, you saw the full spectrum of instructional quality. The continuum of high-, medium-, and low-quality teaching was almost a normal, bell-shaped curve, but this didn't come out in the surveys at all.

Fitzpatrick: Let's come back to a discussion of your use of this model for a minute. Then, I'd like to return to some of the issues you just mentioned. You noted that group cohesion and trust are two important prerequisites for successful use of the empowerment model that were not present in this situation. What other things did you learn from this evaluation about when the empowerment model might be appropriate and when it might not?

Fetterman: The bottom line is "What's the purpose of the evaluation?" If it is for developing and improving the program, then empowerment and participatory approaches are most useful. Traditional approaches *can* help, but they are less likely to. In those contexts where the focus is more strictly on accountability, more traditional modes are useful. Where the purpose is primarily for knowledge development, meta-analysis and other methods are useful, though meta-analysis can be problematic because it's so hard to meet the assumptions. But the first question is "What's the purpose?" The answer to that question helps determine which way to go. Another level to consider is "What do the stakeholders want?" Maybe what they want is strict external accountability. I'm conducting a project right now that is primarily an empowerment evaluation; however, the stakeholder also has a second purpose—external accountability. There are some areas where I have used empowerment very effectively when the focus is external accountability. So these distinctions,

although quite useful, periodically break down in practice. I do not think contrasting approaches like traditional evaluation and empowerment evaluation are mutually exclusive. I would want my bank to use empowerment methods to ask me what hours they should have, what services they should offer, and the like. On the other hand, I want to know where the money is, too. For that, I want an external audit.

A final, but important, issue is whether the organization is conducive to empowerment. You can do empowerment in places that are extremely receptive to it and understand the spirit of it, and also in places that are not conducive but want to be open and receptive to the approach. I have gotten a lot more done in empowerment evaluations when people are receptive to it than when they're not. When the place is authoritarian and dictatorial, they may need empowerment more, even though they're less receptive. But I don't devote most of my time to those places because life is short, and the process is too slow in that environment. When I first started with empowerment, I spent time in every domain (both highly receptive and moderately receptive environments) to see how it worked, but then after a time, you want to work in the places where it will move, to work with people who will help refine it.

Fitzpatrick: Tell me more about the need for group cohesion and trust. That could be a tough standard for many organizations to meet.

Fetterman: There doesn't have to be perfect or complete group cohesion. I did an empowerment evaluation recently with a group that, by their own self-assessment, hated each other. It ended up being one of the best things I've ever done. They ended up realizing that they had so much more in common than they would have guessed at the onset of the process. The process helped them see what they cared about and what they didn't care about. At the end of the first workshop, they were already telling me how much they had in common. They had not recognized how cohesive their group really was already, but I did. In contrast, with the Stanford TEP, they didn't have a "there" there.

The more subtle thing is trust. Even if they hate each other, it's still worth exploring whether there is enough trust to engage in an empowerment evaluation. In empowerment evaluations, most program staff members and participants are much more critical than I would be, so the trust often comes shining through.

Fitzpatrick: Let's move now to focus more on the STEP evaluation. This evaluation has received some national attention in the teacher education field. What particular characteristics of the evaluation or its results have received the most attention?

Fetterman: The key finding, the need to agree on a unity of purpose in a program, has brought responses from directors of other teacher education programs across the United States. This is probably influenced by the fact that the light is on Stanford, and Stanford, for better or worse, is often considered the model, but the absence of unity of purpose in a program has received a lot of attention in and of itself. People are reassessing whether they have unity of purpose.

And it's interesting that this absence of a unity of purpose in the program was one of the last things we found. We found the manifestations at first, but then we realized that the reason all these problems were identified was because of the lack of a unity of purpose. There was no common vision. It was the gestalt of the project.

The second issue receiving attention was our findings on the curriculum and the lack of connection between theory and practice. They didn't have any classroom management courses in the fall quarter. The time you really, really need this is in that fall quarter when you're first beginning to teach. Why didn't they have it then? It was not a convenient time for faculty. As a consequence, students were extremely critical of other courses that quarter that had very little to do with practice. The faculty who taught those courses really appreciated learning about our findings in this regard, in retrospect. The student critique was harsher than it might have been because they were looking for the practice focus even in courses that were explicitly theoretical—because it was absent when they needed it the most. The larger issue here is the idea of having more of a flow between theory and practice within the curriculum.

Another issue that has captured people's attention is our recommendation to teach teachers how to conduct research about their own teaching and for faculty to engage in research activities directly associated with the teacher education program. This might seem obvious in a research institution; however, issues associated with status come into play when we are talking about conducting research in a teacher education program.

A recommendation that appears to have received considerable attention involves maintaining contact with alumni. A lot more places are doing that now. At the beginning of the evaluation, we decided that we were going to interview a lot of the alumni from the program. Our first step was to go to the program and ask for contact information on alumni. We learned there was no list. Guess what? The evaluation is done! They don't even maintain a list! However, that would have been the easy way of dealing with the issue, but it would not shed any light on alumni perspectives. So we contacted a few

alumni we knew and built our own list through snowball sampling. By talking to a lot of alums, we learned that beginning teachers need a lot more support than STEP ever realized. That's when teachers are most vulnerable and the time period in which they are most likely to drop out of teaching. This somewhat circuitous path helped us understand the problems of beginning teachers in a clearer light, and in turn, our findings helped other teacher education program faculty and administrators realize that alumni contact was much more significant than they had realized.

Fitzpatrick: You mentioned the president of Stanford commissioned this study. What prompted his action?

Fetterman: A couple of things: In 1994, significant student dissatisfaction was manifested at the graduation ceremonies. Minority graduates complained about the program, suggesting that it was not responding to minority issues. The new Stanford administration wanted to know what was going on. In addition, this president was interested in education. The president was very clear. He wasn't interested in just making money, although there was some concern about the fiscal administration of the program. His primary concern was with the quality of what we were doing. Although the scope of the evaluation was extensive, he was quite generous with his time and with the funding of the evaluation.

Fitzpatrick: I know the new director of the Teacher Education Program became one of your primary stakeholders, implementing many of your recommendations. Did her arrival prompt the study? Was she involved in the planning?

Fetterman: We had already issued the report on the summer program before she came. However, she was pivotal to the use of the evaluation findings and recommendations. She is considered the most prominent and knowledgeable scholar in the field of teacher education. When she said she loved both the summer interim report and the draft of the final report, defensiveness and critiques of the evaluation disappeared. In addition, she implemented over 50 of our recommendations in the first six months of her tenure. Everything was in alignment. The president was in support of the program and, specifically, program changes that needed to be made. The dean supported the efforts to reform the program. The new director had both the credibility and the force of will to implement the evaluation recommendations, and the evaluation findings were credible and ready to be used. If any one of these things had not been in alignment, use would have been diminished, as is evidenced by past evaluations of this program.

Fitzpatrick: You're a faculty member in the school that delivers STEP. So, in essence, this is an internal evaluation. Evaluators often raise concerns about the objectivity of internal evaluations. On the other hand, internal evaluators' knowledge of the program history, staff, culture, and so forth, can be helpful in insuring that appropriate issues are addressed and information is ultimately used. Let's talk a little bit about these issues. First, how did your prior beliefs and knowledge about this program influence the evaluation?

Fetterman: There was an awful lot I didn't know—for example, about the personal histories of folks. I knew the basic structure of the School of Education—that this was a stepchild. STEP is even jokingly referred to as a stepchild! I always knew there were some problems in terms of status; that is, many faculty viewed it as less prestigious than other assignments or affiliations in the School. So, on the one hand, I was more knowledgeable than someone coming in cold and that was important because at a place like Stanford, if they do not think you have some familiarity with the area you don't have any credibility.

But I didn't know a lot about STEP. If I had known more, it may have been helpful, but then I might not have done the evaluation. Anyone else would have been chewed up. I practically was and I've been here a long time. But there were strengths to my being a little bit of an outsider. If I had been a member of the STEP faculty, I might have been too close to it and thus more likely to just tinker around the edges instead of suggesting bold and fundamental changes in the program. The same way that a fish is the last one to discover water, a faculty member in the program might not be able to see what's in front of him or her without the assistance of the entire group or some outside facilitator or coach. At the same time, I might have been better at seeing some of the detailed findings more clearly.

You're right that there can be a downside to being an internal evaluator and independent judgment can be one of them. However, I didn't have any problem with approximating some form of objectivity or independent judgment. I was certainly perceived by some as too critical. But my colleagues wouldn't accept anything that wasn't a quality evaluation. I am in favor of high-quality teacher education programs. That bias made me even more critical when I saw that the program was not operating as intended. Being in favor of something can make you more critical. If you're in favor of a concept, you take it personally if the program isn't working.

At the same time, a lot of the work is not about describing some objective reality. A lot of what we were doing was telling other people's stories throughout the STEP evaluation because we needed to understand what students were experiencing—what their perception of the program was like as they lived it. We spent every moment in the summer with STEP students—from 7 in the morning until noon at the public school where they practiced teaching and in Stanford classes until 7 at night, and then sorting data at night like students doing their homework. You ended up being more accurate by being immersed in the culture. You're much more sensitive to the nuances and realities when you live the life you are evaluating. When put to the test, you're better prepared to confront fundamental program issues because you have a better insight into what people think and believe.

Fitzpatrick: But this seems potentially conflicting. On the one hand, you note you were able to see some of the big issues because you were *not* part of the STEP faculty and, thus, to build on your analogy, you noticed the water. But, on the other hand, you felt that *being* immersed in the day-to-day culture of STEP with students made you better able to deal with the fundamental issues. What made these circumstances differ?

Fetterman: We are talking about different levels of analysis and immersion. I was not invested in how the place was run because I was not a member of the STEP faculty. This allowed me to think out of the box and question things that most folks took for granted. Similarly, I was not a student, but I needed to understand the students' perspectives in order to understand how the program was working or not working. The best way to get that insight is to immerse oneself in the culture as both a participant and an observer to document the insider's perspective of reality. Daily contact and interaction allow you to collect the kind of detailed data required to describe contextual behavior and interpret it meaningfully.

Fitzpatrick: To what extent were key stakeholders (e.g., faculty) able to perceive you as "objective" in your assessment of the program since you were a part of the education faculty? For example, they might have perceived your recommendation for teaching students more about research as self-serving since you are the director of the School's graduate program in policy analysis and evaluation. What kinds of problems did you encounter in their accepting your findings?

Fetterman: The recommendation about research was broad enough that it was palatable to faculty. We are a research institution and there was little

research being conducted in that area so it was hard to argue with that finding. In addition, because research is highly valued in this environment, it was hard to ignore a finding that was that fundamental to the values of the place.

I also recommended that they have more of a link with educational policy. That should have been a problem, but it wasn't. People recognized that STEP was remiss in not linking with policy (my program) and with the principals' program. I think in some ways we stated the obvious, but no one wanted to confront this issue in the past.

I also believe that the issue was not one of objectivity. The issue here is credibility and honesty. I've done a lot of evaluations for the Board of Trustees at Stanford. They know me and view me as a very honest and straightforward professional, so I had a reputation as being straightforward with the aim of trying to help. I think this was more important than objectivity because what they were really looking for was an honest judgment call about the operations.

I did have one criticism of that nature—lack of objectivity. When the draft report was circulated, one faculty member said I had left out a lot and went to the associate dean saying I had an ax to grind. The associate dean then called him on it, and my colleague had to admit that I didn't have an ax to grind. There was no history of animosity or any other problem between us.

In fact, if anything, one colleague could not understand why we did not have an ax to grind given what we had learned. My evaluation team learned that one colleague *had* behaved inappropriately during the evaluation. During the last week of summer school, when we took off from classes to write the report, this faculty member asked the class, "Are any of the evaluators here today?" When he learned there were none, he had the students close the door and told them not to talk to the evaluators. He said if they (the students) said anything negative to the evaluators, it could end the program. Thirty of these students came to tell us about this colleague's behavior that night. We had already written the section about his teaching, and we had no reason to change it—certainly not because of this behavior. We considered it negative but separate from our assessment of his teaching. At the same time, I could have reported this behavior in the evaluation report. However, I thought it was atypical and would have misrepresented the norm that I would characterize as proper behavior in the program. My colleague still wonders why we will not use that against him to this day. He does not understand the evaluation ethic, which is not to misuse what we learn or distort what we know.

I did think there might be some areas where I would have to excuse myself from certain program assessment activities because of a potential

conflict of interest. However, STEP was a separate enough entity from the School that there really wasn't any significant conflict—it was like reviewing a completely separate program or entity. For all intents and purposes, it was a separately operating program that just happened to be part of the School. That separateness, in fact, was what was wrong with STEP. If it was operating the way it should have been, then I would have known more about it and probably had a conflict of interest. They're even in a separate building.

Fitzpatrick: That's helpful in clarifying how separate they were from you. But, then, your recommendations are to merge more with your own school. Couldn't the nature of that recommendation be perceived as a conflict? You're recommending that STEP would do better if it were in your program. That might be an easier recommendation to make if you weren't a member of that program.

Fetterman: I did not want it to be in my program, just to make links with relevant programs in the school, including mine and the prospective principal's program. These were natural links based on the literature and the recommendation simply stated the obvious—if we want our prospective teachers to be effective they need to know about current policy issues ranging from vouchers to systemic reform. They could acquire this kind of knowledge by making a direct link with the policy program. The same applies to fostering a more productive relationship between teachers and administrators when we recommended that the teacher education program link up with the prospective principal's program.

Fitzpatrick: Usually, more evaluation questions arise during the planning stage than resources permit us to examine. How did you prioritize the questions you wanted to answer? What issues emerged from the planning stage as most critical and why?

Fetterman: Appendix A of the first report listed all of the STEP components and issues that people had mentioned. With this, we found a master list or common denominators that everyone could buy in to. But it was so comprehensive, it meant we had to turn a summer effort into a three-year project. Fortunately, we were given more money and more time to address all the issues. The problem then wasn't which issues do we address, but more what needed to be done first. Not everyone agreed, but we did it by the schedule of the year. We evaluated the summer program first because that was the first event or activity in the calendar. We evaluated students' portfolios when those conferences occurred. We evaluated each component as it came along in the year.

Fitzpatrick: You and your staff became really immersed in the summer program. Three of the five-member team were on-site every day of the summer program. This was quite intensive. Why did you choose this strategy? What did you learn from it?

Fetterman: The need for a five-member team was primarily an issue of size and scope. We couldn't handle that many classrooms with just one or two people. We needed a lot of observations. We were compulsive. I wanted to make sure we had every single classroom covered. We didn't sample—we did them all! We did rotations so we could get an inter-rater reliability going, and we shared digital photographs of what we saw later that day to build reliability and confirm our perceptions. We spent the night organizing and cataloging. We did this at the end of each week and the end of the quarter. Most of the analysis was iterative. Coverage was our number one goal—and quality control to make sure we were on target

Being there every day also gave us more data points to work with. We were able to see the same patterns of behavior over time—that is a form of reliability. In addition, there was a lot of face validity to our observations over time. That's what I love about being in there every single day—it really was ethnographic in nature. If you stay for short observations people have mugging and company behavior, but if you stay a long time they can't sustain that false sense for that long. They get used to you and let go of the company behavior and go back to their normal way of operating and behaving.

Fitzpatrick: Did you structure this initial involvement in the summer program? Or were your observations more open-ended? Tell us a little about what the evaluation team did during their time on-site.

Fetterman: In some cases, the interviews and observations were open ended. In most cases, we had specific things we were looking for. We reviewed the literature associated with teacher education programs and internal program documents and then attempted to document whether program intentions were being actualized. For example, was the student given a chance to teach the class? Did the mentor teacher stay in the classroom to observe? (Sometimes they didn't even stay in to observe the student teacher, even though they were responsible for evaluating their teaching.) We also were assessing the quality or quantity of student engagement in the classroom. We wanted to know if the student teachers were actually "in there" with the kids, or did they fade into the woodwork at the earliest stage and remain that way. Are they really not suited for teaching?

Of course, things emerged that were fascinating, reflecting the dynamic nature of teaching. We saw some great things such as stellar teachers with a certain chemistry with students, and some absolutely pathetic things as well, including lecturing about being student-centered for three hours. Overall, I would say that we used some normal protocols and some open-ended observation. The idea was to describe, not to test. However, observation in our case was really a series of hypotheses about how things should work, and then we used observation to test those assumptions or statements about program operations.

Fitzpatrick: I've written about the importance of observing programs at early stages, and describing and retaining what you see. Tell us a little bit about your view on description at this stage.

Fetterman: It's very important at the early phase of the evaluation. Of course, we had a description of the site (the physical layout) and a lot of very basic preliminary information. But this information wasn't at all insightful about what was really going on, about what chemistry existed between the teacher and the Stanford student and the middle school student. You wouldn't get that without observation. You also need to understand the political interaction between the principal and the teachers in summer school. So, description was critical to get a baseline understanding of dynamics, to understand the challenges and the nature of interactions, and to figure out which interactions to *select* for testing. We needed to know the process to make recommendations for improvements.

We got quite involved in order to provide a sufficiently in-depth description of the program. We hung out during master teacher and student discussions or assessments of student teaching. Luckily, we were well received by the local principal and the teachers. They were very enthusiastic about having us there and providing them with feedback. It was ironic that the university faculty was more defensive and less receptive to critique than the high school faculty.

The different reactions we observed between the university and high school faculty reflected the different norms in each culture or environment. In public schools, you get evaluated and get feedback. You think it's odd if someone comes in and doesn't comment on your classroom. So that culture made the teachers comfortable with our feedback. But that's not part of the culture of academe. So the classroom teachers liked our observation. They wanted us there. They wanted to talk to us about the students. They gave us a little award at the end of the summer as a way of recognizing how important we were as

part of their lives on a daily basis throughout the summer. We were all invested in the program, but we had different roles and responsibilities. Our immersion in the program and their reactions helped us develop a very open, sharing relationship which made the observation much more insightful. We gained a clear understanding about the teacher's role in mentoring students on a day-to-day basis because we were right there with them on a daily basis, we weren't just making cameo appearances.

Fitzpatrick: It sounds like these observations were invaluable in giving you a real feel for the summer program. Let me build on a couple of issues you mentioned. You indicated the teachers wanted to talk with you about the students and that you sometimes considered whether individual students were suited for teaching and chimed in during the debriefing process. Can you tell us a bit about that? Did you see that individual assessment as part of your evaluation?

Fetterman: I did not see it as a formal part of the evaluation. However, it was part of reciprocity in the evaluation, and it was a secondary form of data collection. In other words, our job did not involve the assessment of individual student teachers. It did, however, involve our assessment of the master and coordinating teachers' ability to assess the students. Thus, it was useful to hear their views about individual students as data about their abilities to assess student teachers. On another level, there was an expectation that I would share my opinions as a "fellow teacher" and observer. (I also went through a teacher-training program many years ago, so I am quite familiar with the process and the importance of discussing these matters.) However, most of the time we simply listened without arguing or supporting the teachers' assessments or comments. There were two extreme cases that merited some frank discussion, but this required temporarily stepping out of my evaluation role and into my faculty role to make sure the student received adequate counseling and that the master teacher was apprised of potential problems.

Fitzpatrick: Some evaluators argue that this level of immersion can threaten "objectivity." That is, you may become so involved in the details of program delivery that you begin to identify too much with different audiences, such as deliverers or clients. Did you find maintaining "objectivity" or distance to be a problem in reaching your final judgments or recommendations?

Fetterman: My position is that the problem is completely the opposite. You're going to have superficial or misleading information about a place if you're not immersed. The only way to have a good understanding and an accurate assessment, understanding real-world activity right in front of your eyes, is to be immersed. You have to be part of the culture, to help clean up the lab,

to understand the extra work and the pressures of their personal lives. Basically, I don't agree with the assumptions associated with this question. Immersion is the *only* way to get the best-quality data. And, I mean long-term immersion, repeated involvement over time—to begin to see patterns as we discussed earlier. You can't see patterns without long-term immersion. Immersion is actually spending time with folks and getting their view of why they're doing what they're doing instead of assuming it. Our job is to take these insiders' perspectives of reality and apply our social science external focus.

Spradley (1970), an anthropologist who studied tramps in Seattle, showed the importance of immersion. He spent a great deal of time talking to tramps. Judges wanted to throw them in jail so they had a roof over their heads, but Spradley learned the tramps didn't want that. They wanted their freedom. He described the very different world views of judges and tramps, and he was able to do this because he was immersed in their worlds. So immersion doesn't threaten objectivity. It's probably the only thing that will give us real quality data. You can't be immersed in everything, but the more you limit the immersion the less you learn.

There is no real or absolute objectivity. We approximate concepts of this nature. We hold ourselves up as models of it. Science and evaluation have never been neutral. We always bring our own lens. Most people have a very naive idea about what objectivity is. It's a nice concept, but it's not real. If we delude ourselves with it, we're just perpetuating a myth. We are *all* wearing a lens when we observe or judge something. The generic perception of objectivity is useful, but it's nonsense. Triangulation is a useful tool to help us approximate this concept without being a slave to it.

Fitzpatrick: I know you used qualitative and quantitative methods to test working hypotheses and the generalizability of observations. In what areas did the surveys, interviews, and observations converge and validate each other?

Fetterman: The way I was trained, you can't do good qualitative work unless you use qualitative and quantitative methods. You must use a combination of both methods. There isn't a qualitative world and a quantitative world. There is one world. In addition, you can't be a purist or an ideologue; you must use any appropriate tool available. Sometimes, you're mixing these methods to triangulate—to test or rule out rival hypotheses.

We used surveys to get a handle on the generalizability of specific student views about the program. Individual interviews helped us flesh out, or explore, some of the commonly held views. Similarly, classroom observations were invaluable to cross-check individual interviews and survey data.

There are times when we get overly invested in how we represent our data—descriptively or numerically. For example, in our interviews with faculty, we found no convergence on their views of their mission or place. There was no shared conception of a mission. The convergence was the absence of convergence—they shared a lack of unity of purpose. This could be portrayed in a numerical fashion, such as percentages who believed the mission was one thing and percentages of faculty who thought it was something else. The bottom line is that both descriptively and numerically we found the same thing—that the group had very little in common when it came to a vision of the program mission.

As discussed earlier, we also compared survey data with individual interviews and our own classroom observations to obtain a more accurate assessment of classroom teaching. Any single data source alone would have been misleading. There was a lot of quantitative and qualitative mismatching. In some cases, the results converged; in some cases, they didn't. The lack of convergence prompted us to explore more. It forced us to probe further.

Fitzpatrick: When there was a discrepancy, were the qualitative results always better?

Fetterman: Not always, but most of the time. For example, on the issue of minority enrollment, all the verbal feedback suggested things were problematic but under control, but when we looked at the numbers, it was obvious that it was not under control. The way they were reporting the data was misleading. They were lumping Asians into the minority category to look better. The initial quantitative information suggested good minority enrollment. This issue illustrates the constant interplay between qualitative and quantitative results. After observing the students for a bit, our gut instinct told us the numbers didn't match what we saw. That led us to go back and look more carefully at how the numbers were derived. Our intuition combined with observation led us to believe that the figures didn't make sense. A lot of evaluation is instinct.

Fitzpatrick: The STEP evaluation could be characterized as primarily a process study, a type of study that I consider very useful. But in higher education we generally measure process. Did you consider examining the success of the program in achieving desired outcomes, for example, student knowledge and skills or performance on the job?

Fetterman: We looked at things students were expected to do in order to become good teachers, to see if they were doing those things. The superficial indicators—grades and so forth—suggested all was fine. We did survey alumni and their supervisors and got some information on outcomes in that way. That

was how we learned that students were weak on technology. They weren't trained in a certain category sufficiently.

Fitzpatrick: I think interviews and input from supervisors can be invaluable, but you were learning people's *reports* of what they were doing. Those reports in themselves are very useful for learning what they're struggling with and what they're comfortable with, but you didn't actually go out and observe alumni in their classrooms, did you?

Fetterman: Sure—partly to collect relevant data and partly to establish a bond with them before we asked them to participate in focus groups. However, this was not the focus of the effort—the focus was on the quality of teacher training as a "treatment."

Fitzpatrick: You made use of a variety of technological tools in conducting this evaluation. Your reports present many interesting digital photographs of classroom activities, giving the reader a real picture of the setting and characters. You used the Web for surveys and conducted evaluation team videoconferencing on the Internet. Tell us about these approaches. What worked? What didn't?

Fetterman: Yes, we learned a lot about high-tech and how it can be used in evaluation in this study. I'm on the road a lot, and we used videoconferencing very effectively to keep in touch with evaluation staff and discuss what we were learning. On another level, digital photography was very helpful to us in documenting what we saw in the classroom. Evaluation team members would take pictures of student teachers in the classroom, and we would share them over lunch. These pictures help illustrate what we thought we saw—student teachers fading into the woodwork or actually being involved in the process of teaching. The photos helped document our observations in ways that no one could dispute. It was very powerful. You are able to share with your evaluation team members and others precisely what the teacher is doing, what the students are doing. Once you get the camera, the only expense is the floppy disk—no film or developing required. The photographs also helped confirm our feelings at this early stage. They were a reliability check; the team could consider whether they drew the same conclusion from the picture as the observer had. Pictures also were useful when we prepared the final reports. What a difference a color photograph makes! We went for color for the president's report because it makes such a big impact.

Technology also allowed us to put report drafts on secure areas on the Web to get feedback from the faculty and others. We posted four drafts of the summer report to receive feedback.

We used Nudist as the software for sorting data for all verbatim quotations and observations. It was invaluable—to be able to sort at a moment's notice! You used to have to think twice as to whether to sort again with cards. Now, you can test things out quickly, and it helps keep things organized. We made copies for everyone on the team. If they had an idea, they could play around with it.

We also surveyed alumni of the program using a Web-based survey. The alumni went to the URL and filled out the form; the moment they filled it out, it was sorted automatically. I could be anywhere in the country and sort the data on the Web site. The students could play around with it too.

Fitzpatrick: Let's turn to your results for a minute. Which of the results was most surprising to you?

Fetterman: The political and personal issues in this evaluation were very surprising. I was surprised at how political and nasty it could become. I was astonished at how badly many of the faculty had been hurt by their involvement with the program over the past 20 years. Personally, what I learned about a few of my colleagues' behavior in response to the evaluation was surprising. I don't know if you're ever prepared for the fact that some colleague you've known for a long period of time is either absolutely stellar and very supportive or the opposite where you feel a sense of betrayal and sabotage. Definitely, on a personal level that was quite surprising. When someone close to you, whom you trusted, betrays you, it disturbs your whole sense of judgment and, of course, it's personally disconcerting. The other side is uplifting, when people you don't know that well get up and give speeches in support of you and your work, that's an equally powerful personal experience.

In regard to the actual findings, the two surprises were (1) the differences in what we learned by observing faculty teaching as compared with what we learned from the surveys about teaching and then (2) the lack of unity on purpose or mission. It was surprising that that could even be possible.

Fitzpatrick: Having worked in several different university settings, I have to say that I don't find lack of unity among faculty in academe surprising. Faculty are often either disengaged and doing their own thing in classes or they have strong disagreements about purpose.

Fetterman: Well, you're right, but it's surprising not to find unity of purpose in teacher education. Our review of other teacher education programs showed that they did have this, but our own school did not have a firm sense of purpose of what students should be like when they got out and what they

should be doing in the program to achieve that. Lewis and Clark, Trinity, Bank Street, UCLA, and other schools we looked at all have very specific themes for teacher education. When the school is a professional school, it's important to have agreement on the mission. What is the overarching theme? What do we all agree that our teachers should become? What's our philosophy and value system? And then you need to have that reflected in the curriculum and even alumni contacts. It even affects admission issues.

Fitzpatrick: You did review the programs of other schools quite a bit in your evaluation report. Often, evaluations neglect that area. Tell us about your exploration of other programs.

Fetterman: The comparative part was very helpful to constantly look at other programs that were similar. We're supposed to be the stellar institution. To learn whether we're achieving that, we need to look at what other programs are doing. Folks who are running programs often don't have time for literature reviews; they don't do the research on it. I think it's an error not to look at other programs. To inform others of what's going on in the real world is very helpful. We went to teacher education conferences. Through those we were able to talk to these folks and visit some of their programs. We communicated with them by phone and e-mail too, to get their thoughts and learn more about what they were doing. If we really believe in knowledge, we need to pay attention to what others have done.

Fitzpatrick: When did you do this? Was this part of your planning phase?

Fetterman: Most of it was around the middle of the study. We did some exploration of the literature right off the bat. But, then, halfway through the evaluation, we went to a conference. Then, we got into more depth than in the initial collection of information because we knew more about our program and what we wanted to know. This was very intense information. It helped give us a more normative view, to learn what is reasonable and fair to expect of a teacher education program. The hope is that people will also build on our work.

If you make the review of other programs succinct, clear, and relevant, people will realize there is a value added to learning about other programs. We found it and summarized it. People are always excited to learn what others are doing; they just don't have the time to look into it. You don't want to hit them over the head with it. They can make the decisions with the information we provide. If you use it correctly, you'll get tremendous buy-in for your findings. It places your findings in a larger, more normative context.

Fitzpatrick: I'd like to ask a few questions about your reporting style. Your reports (with the exception of the *Summer School Report*) focus primarily on your recommendations. In this way, you depart from the traditional evaluation format of presenting data to back up your conclusions. Why did you choose this strategy?

Fetterman: I think what you'll find is that in all three reports I'll have the data and recommendations. Of course, the *President's Report* is the shortest; they just wanted a couple of pages, and this has the least data. We wrote about 21 pages, and that is really the maximum for that sort of report. The Executive Summary within that report is two pages long and that is the maximum we thought appropriate for an executive summary.

The STEP report to the faculty and the general community was much more detailed, providing data as needed without cluttering the key points. At the same time, we limited the amount of data in the report for fear of data overload—which happened with our *Summer School Report.* (We still had to have the data and in a format ready to use in case anyone took issue with a specific point in the report.)

The *Summer School Report* was the most detailed because it was the earliest report, the most formative in nature, and the one that required the most feedback to make sure we were starting the evaluation on solid ground. After completing that report, we consciously decided to cut back the length and the amount of data in each subsequent report. I think the detail of the summer report scared the heck out of them. The dean loved it. The teacher education program staff members loved it because it focused on the program detail they had to deal with on a day-to-day basis. But the report was a tremendous amount to digest and in some ways got in the way of the key points or judgments.

My colleagues, the School of Education faculty, focused on the rating of faculty teaching. They would argue with us about our draft memorandum focusing on our observations about their teaching. So we would say, "Here's what we didn't include. Students could barely get in the room because the professor's ego was so big." We thought it appropriate to share our observations of their teaching with them in a draft memorandum format before drafting something about it in the interim report. It gave them a chance to respond to our observations and preliminary findings before it reached the draft report stage. This was ethically sound but personally stressful. We had student comments that were pretty negative, documenting problems with teaching that went beyond what we reported anywhere except to that individual faculty

member. Writing the report was labor intensive, but providing all that information made it more accurate and harder to argue with.

We ended up writing the same kind of detail in the final phase, but just for us. We had to have a tremendous amount of documentation because the evaluation was so political. But we definitely made the decision for a less detailed final report based on the reaction to the first one. The president wanted it brief.

Fitzpatrick: Did you think the 20-page report was a little long for the president?

Fetterman: Yes, but given the nature of everything around it, the political nature of it and the amount of power associated with it, it was too risky to do something conventional like a one- or two-page executive summary without the detail we provided. I wanted to make sure the president had the straight scoop from me because of how others were trying to reach him through other pathways. I negotiated a length of a maximum of 25 pages with him. Because of his academic background, he might not have respected a report of the usual one or two pages. At the same time, the executive summary within this report to the president was only two pages as per the normal custom in evaluation.

Fitzpatrick: That's helpful: I have a better understanding of why you made the final report less detailed than the summer report which did provide summary tables of student ratings and a narrative description of observations in high-, medium-, and low-rated classes. As an aside, I notice you didn't name the instructors, but instead used names of historic figures in education, such as Maria Montessori (who did well) and Edward Thorndike and B. F. Skinner (who were rated low). The descriptions seem to correspond to how one might think these individuals would, in fact, teach. Were these real descriptions of classes, and you just used other names for anonymity? If so, was it disconcerting to others that the descriptions matched how these people might have taught? Did people think it was fictional or real?

Fetterman: The descriptions were real but we wanted to protect the identity of the individual instructors. The pseudonyms were useful devices to both protect the identity of colleagues and signal to the reader almost immediately what kind of teacher we observed.

Fitzpatrick: Let's move into another area. You mentioned to me that you learned a lot about the politics of evaluation and courage from conducting this evaluation. Tell us more about that.

Fetterman: Well, the evaluation may be viewed as wonderful now, but 18 months ago, I was in big trouble. I don't know if I ever completely appreciated

the concept of courage until things hit the fan in a more personal way. It's hard when it's with people you've known for a long time. I usually don't get too caught up with this concern about courage. Chelimsky and Scriven have spoken about it. However, I have not given it a lot of thought over the years. However, this was just unusually personal and nasty on the negative side. There were real highs and real lows associated with the conduct of this evaluation. It was even discussed at my performance appraisal. It's easy to be retrospectively courageous. The hard thing was to do the right thing at the time.

Fitzpatrick: Could you give us an example?

Fetterman: Just continuing the assessment when you're hit with some defensive and inappropriate kinds of behavior becomes a task—just continuing to report negative findings, rather than minimizing them, is a challenge. We could have written the report in such a manner as to make the program look better than it deserved, and it would have helped my own career. In other words, it would have been easier not to take the flak. However, I am glad we stuck with it and continued to provide an honest account of what we observed, even though it was not always pleasant. It certainly paid off in the long run. One example was giving individual faculty feedback on their teaching in the initial draft report. By being honest and giving feedback early, we allowed defensive and periodically combative faculty behavior to emerge in the short run, but this approach minimized conflict in the long run. In other words, in the short run this open, sharing approach did contribute to game playing; the early feedback gave them a chance to attack. I knew at that point that without question we needed to have absolutely solid data and a clear-cut chain of reasoning. If it were too general, it would sound like loose and sloppy methodology, intuitive impressions without the solid backing required to accompany it. It would have been much easier not to report back, but I wanted to make sure that it was accurate at the individual level. So it took courage to go to my colleagues and say, "This is what we found." It took courage dealing with the dean—your boss—knowing how the troops are reacting to him and the report. There was a concern that the evaluation might present a risk to the entire school if these kinds of results went to the president. Politically, conducting the study was a big risk for me—having to be honest in a critical way with people you will have a continuing relationship with. It took courage to keep advocating for students who will be gone while I remain with colleagues I may have alienated.

Fitzpatrick: Do you think people realized there would be so much focus on their own teaching?

Fetterman: Yes, but I don't think they ever expected the detail. But you can't be general without individual data. I needed the individual data for the evaluation to be persuasive. However, I think some faculty seriously didn't know where they stood. I think those at the very bottom and top knew where they were. The people in the middle area were the most significant problem; they thought they were stellar. They thought the evaluation of their teaching was a personal attack. They were so used to subterfuge and combat that they weren't able to initially understand the data. Because it was direct, simple, and straightforward, it was counter to the culture.

Fitzpatrick: But focusing on their individual teaching is more like performance appraisal or personnel evaluation than program evaluation. Evaluation often focuses more on the program than the individual. How did your focus on individual teaching come about?

Fetterman: The needs assessment for the evaluation highlighted the need to focus on people, on individual faculty teaching. We asked what everyone recommended that we look at in the evaluation. They thought a focus on teaching was important as well. We did a six-month planning phase focusing on what key stakeholders thought were the most important things to look at. We decided, based on the time we had, what we would do realistically. However, after consulting with the president about the scope, he simply gave us more resources to do it all. The faculty had their input; they just didn't expect the evaluation of teaching to be anywhere near as thorough and detailed as it was.

Fitzpatrick: Did you encounter any ethical challenges in this process?

Fetterman: Oh, yes, constantly. There was the ethical dilemma of how to handle the information about the faculty member who told the students not to talk to us. I could have reported this to the president, but decided not to. I did let cognizant individuals within the department know, but reporting this to the president would, I felt, affect the findings of the evaluation. To this day, my colleague doesn't understand why I haven't used this information against him. He considers me odd. To him, it's a political decision. To me, it's an ethical decision not to use the evaluation data against him; it's not ethical to use the evaluation for personal reasons or to demean someone. I think he's thinking I'll use it to get something from him. I don't believe in doing that. To do this, professionally, would demean the field.

Fitzpatrick: But you did report it to someone in the department. Did you see that as using the information against him? What was the impetus for that

action? Since his actions didn't seem to inhibit the students from coming to you, they didn't seem to have an impact on the evaluation.

Fetterman: Two things: (1) On a micro level, it was necessary to safeguard the evaluation effort by documenting the event and reporting it on a local level to prevent his behavior from disrupting our efforts or having a chilling effect on communication for any segment of the student population and (2) just because a good segment of the student class came to us does not mean it did not have a chilling effect on those who did not come to us. This kind of behavior does undermine an evaluative effort. However, if taken care of, it does not have to become part of the evaluation report.

Fitzpatrick: Were there people pushing you to make the program look more positive than it was?

Fetterman: Quite frankly, I was so busy getting the job done, if someone hinted at that I didn't listen on that level. The challenges were tremendous. The evaluation experience was not all negative. It was exhilarating! We had something that we all cared about and the study had implications for all of us. We all knew it was important to do. We had a personal connection to the students. We felt an obligation to help them out. It was exhilarating working with such talented faculty and talented students.

Fitzpatrick: We all learn from evaluations. If you were to conduct this evaluation again, what would you do differently?

Fetterman: I would do an empowerment evaluation! I think I would! I'm half serious. I obviously made an assessment that empowerment wasn't the approach for that group at that time. But now they're ready. It's easier to be honest with them now; there's so much that's on the table already. The key point now is not so much what I would do differently in that evaluation, but what I would do now for the next step. They're ready for empowerment evaluation. My role would be as coach and facilitator.

Fitzpatrick: Do you think it would be difficult to step into that role now having had the political turmoil that arose in this evaluation and with your having played a quite different role? Would you be able to establish a trusting relationship with them?

Fetterman: Yes. In fact, I am already beginning a much more participatory or empowerment-oriented evaluation focusing on the feasibility of developing an elementary teacher education program at Stanford. This will probably require an empowerment evaluation of the new program, which may at one point lead to an empowerment evaluation of the older secondary school program. Time and distance make a big difference.

Fitzpatrick: David, thank you for sharing the details of this evaluation with us. As practicing evaluators, we learn from hearing the experiences, difficulties, and choices faced by others. It informs our own practice.

Fitzpatrick's Commentary

This interview helps us learn more about the important choice of models or approaches to use when conducting an evaluation. Although Fetterman is recognized for empowerment evaluation, he rightly acknowledges, even emphasizes, that evaluators should choose a model that is appropriate to the context and purpose of the evaluation. To use the same model in all settings would be unwise and inappropriate. Thus, while Fetterman often uses an empowerment approach, he also has used other approaches. Here, he incorporates elements of empowerment but ultimately takes a decision-making approach focusing on the president of the university as the primary audience. He also makes use of elements of an older model, discrepancy evaluation (Provus, 1971), by studying the process of the program to determine if it conforms to the intended model and to other exemplary models of teacher education. The importance of a trusting environment and culture, which he perceived to be lacking at the time of his study, appears central to a successful empowerment approach. Their absence and the information needs of the key stakeholder prompt him to pursue a different model.

This interview is also enlightening in helping us learn about the strengths and weaknesses of the internal evaluator's role. First, the interview reflects the difficulty in categorizing an evaluator as purely internal or purely external. In Fetterman's case, he would be considered an internal evaluator because he was an employee of the same organization as STEP, Stanford University. Although Fetterman did not teach in STEP, he was a faculty member in the College of Education, and as he indicates, the evaluation required him to judge a program conducted by colleagues with whom he had, and would continue to have, relations for years to come. Conversely, he might be considered an external evaluator because he is not a member of the STEP faculty; that is, he was not employed in the unit to be evaluated. As such, he had some distance and independence from STEP. The head of STEP, for example, was not his direct supervisor. He did, however, report to the same dean. That closeness was illustrated by the fact that a discussion of the evaluation arose at his annual performance appraisal. Another element that is important in characterizing his role in this evaluation is that he did not typically play the role of an evaluator

at Stanford. Unlike many internal evaluators, that was not a permanent part of his duties. Instead, he was a tenured faculty member with the responsibilities that that position entails. He was thrust into a different role here, though he indicates he has conducted previous evaluations for Stanford. Fitzpatrick, Sanders, and Worthen (2004) and Mathison (1999) describe the internal-external position as a continuum rather than a discrete distinction, and that continuum is evidenced here.

Fetterman's position illustrates some of the strengths and weaknesses of an internal role. His knowledge of Stanford University and its culture helped him in considering the types of evidence that would be valid to the president and to other stakeholders within the university. His familiarity with the environment of the organization allowed him to be accepted and gave him knowledge of the important actors and the manner in which decisions are made. And his continued presence in the organization permitted him to continue to encourage use of the results. In fact, because his recommendations included closer relationships between his department and STEP, he was one of the people responsible for implementing these recommendations. But this strength also hints at the weaknesses of the internal evaluator. Does an internal evaluator have sufficient independence to judge the program? To provide a new perspective? For others to trust that his conclusions are "unbiased"? Fetterman believes that he was able to bring an objective, independent judgment to the evaluation, but the turmoil surrounding the results suggests that some his colleagues, correctly or incorrectly, disagreed. Their views may have been inflamed by the fact that the evaluation described and judged the teaching performances of individual faculty who, though anonymous, may have been identifiable by description. Similar faculty reactions may have emerged in response to these descriptions of individual's teaching performance whether one was an internal or external evaluator. But the concerns about the independence of an internal evaluator are most evident when findings or recommendations have perceived advantages for the individual evaluator, however valid the findings or recommendations may be. Thus, Fetterman's recommendations regarding work with his own department and the need for prospective teachers to learn more about research fall into this category.

The evaluation of STEP is ultimately a process evaluation providing information for formative purposes. Fetterman indicates that the president might use the evaluation for summative decisions (to decide whether to continue the program), but the results appear to be sufficiently positive to merit continuation and then were used for program improvement. As such, Fetterman's

methods—immersion in the summer program; extensive interviews with faculty, students, and alumni; reviews of other programs; surveys of alumni and supervisors for feedback—provided a foundation for a view of how the program was delivered and for the elements that appeared to conform to expectations and those that did not. The use of multiple measures (interviews, surveys, observations) and multiple audiences (students, faculty, teachers in the field) added to the wealth of information provided. Finally, his use of technology—digital photography to confirm and demonstrate observations, surveys on the Web, posting of draft reports for the evaluation staff and faculty—facilitated effective communication among stakeholders in a cost-effective and efficient way.

The evaluation raises a number of issues to consider: the choice of an evaluation model and the focus of an evaluation; the politics faced by internal evaluators; the utility of immersion in gaining an authentic sense of what happens in a program; the overlaps that can occur in evaluation in making assessments of clients (students), staff (faculty), and the program itself; and the choices one faces in depth and style of reporting.

DISCUSSION QUESTIONS

1. What is the purpose of Fetterman's evaluation?

2. Description is a key part of Fetterman's evaluation. What does he describe? What elements is he most concerned with? How does description fit with the purpose of his evaluation?

3. Discuss the methods Fetterman uses to describe STEP. In particular, consider his use of immersion into the summer STEP program. How does immersion help his evaluation?

4. Many evaluators whom we interview make use of observation. Contrast Fetterman's immersion and his role in it with Greene's observation of the leadership program. How are their methods and roles alike? How are they different? In a later interview, Ross Conner's staff observe community meetings. If you have already read that interview, consider how Conner's observation procedures differ from Fetterman's immersion and how they are alike.

5. Would you characterize Fetterman as more of an internal evaluator or an external evaluator? In this particular evaluation, what are the strengths of his

being internal, in the sense of being a member of the education faculty at Stanford University? What are the drawbacks, not only to him but to the evaluation?

6. Fetterman's emphasis is more on description than on outcomes. Do you agree with his choice? What factors may have influenced his choice?

7. What outcomes might you choose to measure if you were evaluating this program? Consider both short-term and long-term outcomes. How might you measure them?

8. How does Fetterman involve other stakeholders in his evaluation? Would you consider this a participatory evaluation?

FURTHER READING

Evaluation reports on the STEP evaluation are available at www.stanford.edu/davidf/step.html.

Fetterman, D. M., Connors, W., Dunlap, K., Brower, G., & Matos, T. (1998). *Stanford Teacher Education Program 1997–98 evaluation report.* Stanford, CA: School of Education, Stanford University.

Fetterman, D. M., Connors, W., Dunlap, K., Brower, G., Matos, T., & Paik, S. (1999). *A report to the President: Stanford Teacher Education Program 1997–98 evaluation.* Stanford, CA: Stanford University.

Fetterman, D. M., Dunlap, K., Greenfield, A., & Yoo, J. (1997). *Stanford Teacher Education Program 1997 summer school evaluation report.* Stanford, CA: Stanford University.

REFERENCES

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2004). *Program evaluation: Alternative approaches and practical guidelines* (3rd ed.). Boston: Longman.

Mathison, S. (1999). Rights, responsibilities, and duties: A comparison of ethics for internal and external evaluators. In J. L. Fitzpatrick & M. Morris (Eds.), *Current and emerging ethical challenges in evaluation* (New Directions for Evaluation, No. 82, pp. 25–34). San Francisco: Jossey-Bass.

Provus, M. M. (1971). *Discrepancy evaluation.* Berkeley, CA: McCutchan.

Spradley, J. P. (1970). *You owe yourself a drunk: An ethnography of urban nomads.* Boston: Little, Brown.