Journal of Curriculum and Supervision
Spring 1989, Vol. 4, No 3, 211–229



# MEASUREMENT VERSUS SUPERVISORY JUDGMENT: THE CASE OF *SWEENEY V. TURLINGTON*

HELEN M. HAZI, *West Virginia University*

In the 1984 school year, Maryanne Sweeney, a veteran teacher, applied for the title of Associate Master Teacher in Florida's Master Teacher Program and was denied. Although she scored at the 86th percentile on the subject area exam, she scored at the 18th percentile on an instrument that measured her classroom teaching performance, the Florida Performance Measurement System (FPMS). Her score on the FPMS disqualified her from the title and a $3,000 incentive award.

When she challenged the reliability of the FPMS in an administrative hearing, the decision was upheld. She had an effective lesson and was well prepared, in the opinion of her observers, but this opinion was of no consequence because her observers were merely data collectors for the system and were not to render judgment.

In a recent article, Sergiovanni describes the confusion between measurement and evaluation and how measurement-oriented evaluation systems diminish the role of the evaluator.[1] The case of *Sweeney v Turlington* dramatically illustrates this notion.[2] It was 1 of 200 cases filed by Florida teacher associations against the state in the first year of the Master Teacher Program. This story needs to be reconstructed and told so others can understand its lessons on the practice of teacher evaluation and supervision, especially since at least 11 other states use instruments like the FPMS for large-scale assessment of teacher performance for initial certification, career ladders, or annual evaluation.[3] Using three volumes of hearing transcripts, legal documents (e.g., depositions, statutes), educational writings, and telephone interviews, I first

[1]Thomas J Sergiovanni, "Will We Ever Have a True Profession?" *Educational Leadership* 44 (May 1987) 44–49.

[2]*Sweeney v Turlington and the State Board of Education,* Final Order, Case No 86-0023, Department of Education, Tallahassee, FL (1986, September 22)

[3]Chad Ellet, Emerging Teacher Performance Assessment Practices Implications for the Instructional Supervision Role of School Principals, in *Instructional Leadership Concepts, Issues, and Controversies,* ed. W Greenfield (Boston: Allyn & Bacon, 1987), pp 302–327.

reconstruct, then interpret the case by identifying its salient themes and lessons.[4]

## FLORIDA: THE REFORM CONTEXT

Florida has been characterized as one of the five "bellwether" states in the country where new trends develop and as a "high change" state with a history of reform.[5] When *A Nation at Risk* called on the states to upgrade student achievement and teacher quality, Florida had already been involved with legislation and litigation concerning student-competency testing since the late 1970s.[6]

Reform was not new to Robert Graham, the "Education Governor." Graham had a long-time interest in education as chair of the state's Senate Education Committee. As governor, he wanted to be the first in the nation to come out with a merit-pay plan. In fact, some believed he was in a race with Tennessee's Governor Lamar Alexander for such a plan.

In 1982, his Commission on Secondary Education produced a report that became the priority agenda of the 1983 legislature. A broad reform initiative, financed primarily by the business community, resulted in increased gradu ation requirements, salary increases, examination of beginning teachers, and principal academies. To ensure the continued participation of the business community (and their money), the legislature was being pressured to adopt a merit-pay plan.

Since the legislature was unable to address merit pay by the end of the 1983 session, a special interim commission was formed, the Florida Quality Instruction Incentives Council, or "QIIC" (pronounced "quick"). QIIC provided legislation for both school-based and individual merit pay. The school based program, known as the District Quality Instruction Incentives Program, provided funding for districts to develop and carry out their own merit pay programs. The other program, known as the State Master Teacher Program, provided the legal context for *Sweeney* to evolve.

The intent of legislation for the Master Teacher Program was "to recognize superior ability among instructional personnel . . . and to provide an economic

---

[4]The purpose of this research was to retroductively reconstruct the events, reveal the reasoning in, and interpret *Sweeney*. The reconstruction came from analyses of case transcripts and documents and interviews with its key participants A case study protocol according to Yin provided the guide for conducting the research. See Helen M. Hazi, "Measurement v Supervisory Judgment. The Case of *Sweeney v Turlington*" (paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 1988) See also Robert K. Yin, *Case Study Research. Design and Methods* (Beverly Hills, CA. Sage Publications, 1984).

[5]See John Naisbitt, *Megatrends* (New York. Warner, 1982) The description of reform in this section comes from William Chance, *The Best of Educations. Reforming America's Public Schools in the 1980s* (Denver. Education Commission of the States, Catherine T. MacArthur Foundation, 1986), pp. 78–88.

[6]The notable *Debra P v Turlington*, 564 F Supp 177 (M D FL 1983), case illustrates Florida's early involvement in reform.

incentive to such personnel to continue in public school instruction."[7] To qualify, teachers had to have at least four years' teaching experience (two years in state), a master's degree in their field, a superior score on a subject-area exam, and a score at or above the 75th percentile on a performance evaluation conducted by their principal. The principal had to use "a reliable, valid, and normed performance-evaluation system approved by the State Board of Education."[8] The State Department of Education gave counties a choice of using the FPMS or developing their own reliable, valid, and normed performance-evaluation system. All but one chose the FPMS.

## THE FLORIDA PERFORMANCE MEASUREMENT SYSTEM

The novel feature of this merit-pay plan was its use of performance assessment with inservice teachers. At this time, states were just legislating performance assessment to certify beginning teachers. Today, Florida is 1 of at least 12 states to mandate large-scale assessment of teacher performance for initial certification, career ladders, or annual evaluation.[9]

The FPMS is a performance-assessment system of two instruments. one summative, which is used to screen teachers to identify problem areas and to compare teachers, the other, formative, which is used to pinpoint behaviors for remediation. It was originally developed by B. O. Smith and Donovan Peterson at the University of South Florida for the Florida Beginning Teacher Program.[10] Observers used the FPMS Summative Observation Instrument in the Master Teacher Program to evaluate Maryanne Sweeney.

The FPMS is a way to "[cut] the knowledge base of teaching into manage-able bites."[11] The FPMS contains 6 "domains," subdivided into 34 "concepts," which consist of 124 "behavioral indicators." On the basis of their review and analysis of process-product and experimental research on teaching, the devel opers included such items as "uses body behavior that shows interest—smiles, gestures", "maintains instructional momentum", "discusses cause-effect/uses linking words/applies law or principle."[12]

The intent of the Summative Observation Instrument is to measure 21 in-classroom behaviors from 4 of the 6 domains. The developers omitted

---

[7]"State Master Teacher Program," *Florida Statutes*, sec. 231 533 (1983)

[8]Ibid., sec 231.533(c)1.

[9]Chad Ellet, 'Emerging Teacher Performance Assessment Practices. Implications for the Instructional Supervision Role of School Principals," in *Instructional Leadership. Concepts, Issues, and Controversies*, ed. W. Greenfield (Boston. Allyn & Bacon, 1987), p 304

[10]It has been labeled a forerunner of state beginning-teacher assessment programs See Linda Darling-Hammond and Barnett Berry, *The Evolution of Teacher Policy* (Washington, DC Center for Policy Research in Education and Center for the Study of Teaching, Rand Corporation, 1988)

[11]C J B. Macmillan and Shirley Pendlebury, 'The Florida Performance Measurement System A Consideration," *Teachers College Record* 87 (Fall 1985). 70.

[12]B O. Smith, Donovan Peterson, and Theodore Micceri, "Evaluation and Professional Improvement Aspects of the Florida Performance Measurement System," *Educational Leadership* 44 (April 1987): 16–19.

indicators for instructional planning and evaluation of student progress because they believe that such behaviors mainly occur outside the classroom and thus cannot be observed. The developers claim that the FPMS meets both "legal and professional requirements."[13] It has been normed on 1,223 beginning teachers and 20,000 master teachers, and its content has been validated by national experts.[14] The FPMS meets the legislative mandate for a reliable, valid, and normed performance-evaluation system and has withstood legal challenge since *Sweeney*.[15]

Since 1984, 10,000 individuals have been trained to use the FPMS; 4,500 have been principals and assistant principals; the remaining have been district-level staff members, classroom teachers, university personnel, and department of education personnel.[16] To be a certified observer, an individual must attend three days of training, pass content exams on teacher-effectiveness research and the coding manual, and successfully code videotaped lessons using the FPMS.

The FPMS is no stranger to description or scrutiny and has been a subject of print since 1973.[17] Some call it "the most extensive attempt in recent years to translate research on teaching into a practical form for use in training, evaluating, and rewarding teachers."[18] Its writings can be classified as description,[19] research report,[20] or criticism.[21] It receives mention in status pieces[22] and in commentaries[23] on educational policy and reform.

---

[13]Ibid.

[14]Experts included Nathaniel Gage, David Berliner, Donald Medley, and Joe Mazur

[15]Donovan Peterson and B O Smith, "The Use of Teacher Effectiveness Research in the Preparation of Instructional Supervisors," *Wingspan The Pedamorphosis Communique* 3 (December 1986): 15–19.

[16]In addition, those outside of Florida trained in the use of the FPMS include 200 observers from nearby islands, 200 from the state of Washington, 5,000 from Kentucky, and some from Colorado. See Arthur Shapiro and Marion Romens, "Interview with B O Smith and Donovan Peterson," *Florida ASCD Journal* 4 (Fall 1987) 24–27 In addition, 40,000 copies of the FPMS have been sold throughout the country, and individuals have consulted with Oklahoma, Texas, Maryland, Alabama, Tennessee, and North Carolina at the request of their state departments

[17]In fact, the FPMS is an identifier in ERIC, a tentative status given to proper names or concepts to see whether they should become permanent descriptors in the *ERIC Thesaurus* A computerized ERIC search yielded 8 articles in *Current Index of Journals in Education* and 13 entries in *Research in Education*

[18]C J B Macmillan and Shirley Pendlebury, "The Florida Performance Measurement System A Consideration," *Teachers College Record* 87 (Fall 1985). 67.

[19]See, for example, Thomas H Fisher, Betty V Fry, Kenneth L. Loewe, and Garfield W Wilson, "Testing Teachers for Merit Pay Purposes in Florida," *Educational Measurement Issues and Practice* 4 (Fall 1985) 10–12, B O Smith, Donovan Peterson, and Theodore Micceri, "Evaluation and Professional Improvement Aspects of the Florida Performance Measurement System," *Educational Leadership* 44 (April 1987) 16–19; Donovan Peterson, Jeffrey Kromrey, Theodore Micceri, and B O Smith, "Florida Performance Measurement System An Example of Its Application," *Journal of Educational Research* 80 (January/February 1987) 141–147

[20]See, for example, University of South Florida, *Teacher Evaluation and Assessment Center Report for 1984–85* (Tampa University of South Florida, College of Education, 1985), University of South Florida, *Teacher Evaluation Study Beginning Teacher Portfolio Study, FPMS Validity*

## BACKGROUND TO THE CASE

The setting of the events leading up to the case was Clearwater High School in Pinellas County, Florida. During 1984–85, the high school had approximately 2,500 students, 100 teachers, and 6 assistant principals The case had 6 key participants.[24] Maryanne Sweeney was a computer teacher and chair of the Business Education Department with 15 years of teaching experience. Three observers were witnesses for her defense: Bill Williamson, principal, in his 10th year; John Nicely, in his 2nd year as an assistant principal; and Dorothy Cheatham, an assistant principal brought in to re-evaluate Maryanne. FPMS developers who were witnesses for the state included Betty Fry, the official spokesperson for the State Department of Education; and Donovan Peterson, senior researcher, co-developer of the FPMS, and a professor in the Department of Leadership, University of South Florida.[25]

In the fall of 1984, Bill Williamson conducted a meeting with the teachers who voluntarily applied for the Master Teacher Program. Maryanne, having passed the subject-area exam by scoring at the 86th percentile, was 1 of 15 teachers present. At the meeting, Williamson gave the teachers a copy of the instrument, but as instructed, did not discuss it.

On February 20, 1985, Williamson observed Maryanne teaching 22 students in an Introduction to Data Processing class. The room had five operating terminals. The purposes of the lesson, in the words of the teacher, were to "handle the topic of updating files through editing, ... collect the homework. ., make a couple of reviews, teach the commands, and then have students practice at several different types of activities."[26]

Williamson believed: "It was a good lesson ... I felt it was a very fine lesson. ... I was constrained not to discuss it with her, but I felt she had done

---

*Studies, Master Teacher Score Trends, Special Instruments Norming* (Tampa. University of South Florida, College of Education, 1986).

[21]See, for example, C. J B. Macmillan and Shirley Pendlebury, "The Florida Performance Measurement System· A Consideration," *Teachers College Record* 87 (Fall 1985). 68–78.

[22]See, for example, Chad Ellet, Emerging Teacher Performance Assessment Practices Implications for the Instructional Supervision Role of School Principals," in *Instructional Leadership. Concepts, Issues, and Controversies,* ed W. Greenfield (Boston. Allyn & Bacon, 1987), pp. 302–327, Gary R. Galluzo, Assessment of the Teaching Skills of Beginning Teachers," in *What's Happening in Teacher Testing: An Analysis of State Teacher Testing Practices,* ed. L. M. Rudner (Washington, DC Superintendent of Documents, U S. Government Printing Office, 1987), pp. 39–42.

[23]See, for example, Linda Darling-Hammond and Barnett Berry, *Evolution of Teacher Policy*, (Washington, DC. Center for Policy Research in Education and Center for the Study of the Teaching Profession, Rand Corporation, 1988); Thomas B. Timar and David L. Kirp, "Education Reform and Institutional Competence, ' *Harvard Educational Review* 57 (August 1987). 308–330.

[24]All names have been retained because the case is a matter of public record

[25]Peterson teaches courses in the supervision of instruction and in teacher evaluation, is co-director of the Teacher Evaluation and Assessment Center at the University of South Florida, and is founder of a teacher-evaluation consortium in the Southeast.

[26]Testimony from Vol. I, p. 35 of the transcripts of the case, *Proceedings Volume I–III, Sweeney v Turlington and the State Board of Education,* Case No 86-0023, Clearwater, FL (1986, April 15)

a nice job, and it would have been good to tell her that."[27] Even though he thought the lesson was good, he could not discuss the observation with Maryanne. Instead, he asked her to sign the FPMS form so it could be sent to a computer for scoring. He believed, however, that the teacher should receive feedback: "I feel very strongly that when you evaluate someone there should be feedback, other than saying, 'Nice job,' or, 'You're doing well,' or something like, a general comment."[28] In fact, he believed the form he was completing was telling her she had taught a good lesson.

Because two observations were required, John Nicely observed Maryanne on February 21, the next day. She taught the exact same lesson but to a different group of students. Maryanne was the first or second teacher Nicely had observed. Like Williamson, Nicely believed that it was "a very good lesson":

> I felt there was learning taking place on the students' part I thought that the instructor, Ms. Sweeney, was well prepared. She was getting the point across that she was trying to get across. She then gave the students an opportunity to ask questions, to discuss. She also would give the students an opportunity to practice what they had been talking about and discussing in the class during that period. . . . I felt it was good. I enjoyed it, and I felt the students had learned something.[29]

Nicely reaffirmed his judgment about the quality of the lesson and raised the possibility of some error in the score on his part or on the part of the system:

> There was no way, in my opinion, that it was an 18 percent performance that day for her. Either I did something that was wrong, or the instrument didn't measure something it was supposed to measure, was my feeling.[30]

He, too, believed that the FPMS was telling Maryanne that she had taught a good lesson "My opinion is that it was a good class and that I had marked the instrument the way I saw things happen."[31]

After the two observations, Maryanne was notified that her forms had been received and that she had scored below the 50th percentile. She then requested her exact score in writing and a re-evaluation. Teachers may request a re-evaluation if it is mathematically possible that they can obtain a higher score.

On August 29, 1985, the beginning of the next school year, Dorothy Cheatham was brought in for the re-evaluation. Even Cheatham believed that she observed "an effective lesson." Her testimony confirmed what the others believed—that the score did not accurately reflect the quality of lesson they had observed.

Even with the third evaluation and 10 bonus points from her principal, Maryanne scored only at the 18th percentile and thus did not qualify for

---

[27]Ibid., p 129
[28]Ibid.
[29]Ibid., p. 108.
[30]Ibid., pp. 108–109
[31]Ibid., p. 110.

Associate Master Teacher and a $3,000 bonus. She appealed the decision to the Commissioner of Education, the administrative case became known as *Sweeney v. Turlington.*

## THE ARGUMENTS

In his opening argument, Ron Myers, Maryanne's attorney, captured the essence of the case when he compared the FPMS to "police radar" and argued that radar may be a valid tool, "but if the radar gun isn't properly calibrated at the beginning, if the operator of it isn't properly trained, we run into . . . the trooper clocking a tree at 73 miles an hour."[32] Throughout the case, Meyers argued that Maryanne did, in fact, exhibit behaviors that should have qualified her as an Associate Master Teacher. He claimed that her observers "underrated" a superior-performing teacher, challenging the reliability of the instrument, the training of observers, and their accurate completion of the instrument.

In defending the FPMS for the state, the lawyer presented witnesses who testified on how and why the FPMS was developed, why it is reliable, why it meets all the standards set out by the statute and rule, how the FPMS differed from annual ratings, and that "it fairly and reliably assessed [Maryanne's] classroom performance."[33]

Maryanne and her lawyer had to present a legal, not an educational, argument; they had to find some statutory or regulatory grounds that the FPMS violated or did not fulfill. They had the burden to show that the FPMS "did not [achieve] the purpose of the statute . . . to recognize superior ability among instructional personnel."[34] Unfortunately, Meyers took issue with the use, not the validity, of the instrument.

## THE THEMES

Breaking from the narrative and moving to the interpretive mode, themes are next presented to use further details about the case to explicate its significance for the field of supervision. These salient themes were identified once the story was reconstructed. A theme may have been sparked by a witness's quote, an inventive term and its accompanying definition, or an unchallenged claim embedded in an argument. Identifying themes is a step toward interpretation—"the search for meaning in the events under study."[35] Themes begin with the FPMS, the center of the controversy, then move to the observers who use it and the teachers who are the objects of it.

---

[32]Ibid., p. 12.
[33]Ibid., p. 14.
[34]Ibid., p 16.
[35]Noreen B. Garman, "The Clinical Approach to Supervision," in *Supervision of Teaching,* ed. Thomas J. Sergiovanni (Alexandria, VA: Association for Supervision and Curriculum Development, 1982), p. 51

### *The Myth of the Objective Instrument*

Certain words were used to characterize the FPMS: *based on teacher-effectiveness research; generic; objective; precise.* This language was an effective defense because each claim went unsubstantiated and unchallenged. The language also contributed to the myth that some objective instrument exists.

*Based on teacher-effectiveness research* was one catch-all phrase used to legally defend the FPMS. Its developers claimed that teacher-effectiveness research represents the consensus of the field about what constitutes good teaching. At one point in the testimony, teacher-effectiveness research became synonomous with the research base of the FPMS, as if they were one in the same.

To the contrary, teacher-effectiveness research does *not* represent a consensus of the field about what constitutes good teaching.[36] Some scholars believe that the knowledge base of teaching has yet to be discovered.[37] They view the scope of this research as limited, as focusing exclusively on teaching skills (as opposed to concepts or complex material) at the elementary level, and as resulting in "findings [that] have been much more closely connected with the management of classrooms than with the subtleties of content pedagogy."[38] Shulman describes how policymakers have erroneously used this research:

> When policymakers have sought "research-based" definitions of good teaching to serve as the basis for teacher tests or systems of classroom observation, the lists of teacher behaviors that had been identified as effective in the empirical research were translated into the desirable competencies for teachers. They became items on tests or on classroom-observation scales. They were accorded legitimacy because they had been "confirmed by research." While the researchers understood the findings to be simplified and incomplete, the policy community accepted them as sufficient for the definitions of standards.[39]

---

[36]Examples of those who have recently summarized this body of research include J. J Brophy and T Good, "Teacher Behavior and Student Achievement," in *Handbook of Research on Teaching,* 3rd ed., ed M C. Wittrock (New York. Macmillan, 1986), pp 328–375, B Rosenshine and R. S Stevens, "Teaching Functions," in *Handbook of Research on Teaching,* 3rd ed., ed M C Wittrock (New York. Macmillan, 1986), pp 376–391 and L Shulman, "Paradigms and Research Programs for the Study of Teaching," in *Handbook of Research on Teaching,* 3rd ed., ed. M C Wittrock (New York Macmillan, 1986), pp 3–36.

[37]See, for example, Lee S Shulman, "Knowledge and Teaching Foundations of the New Reform," *Harvard Educational Review* 57 (February 1987) 1–22

[38]Ibid., p 10 Sergiovanni places the research on generic teaching effectiveness into perspective He says that the research affirms well-established craft knowledge and is therefore a scientific contribution "But its [the research's] scientific validity can only be claimed on a case by-case basis The merit of the research is the rendering of the best conditions for particular teaching behaviors to lead or not lead to increased student achievement. These insights represent important scientific contributions It is scientistic, on the other hand, to ignore contextual constraints and to assume instead that the research is intended to or has revealed universal teaching principles that constitute a best way to practice." See Thomas J Sergiovanni, "Science and Scientism in Supervision," *Journal of Curriculum and Supervision* 4 (Winter 1989). 96

[39]Lee S Shulman, "Knowledge and Teaching Foundations of the New Reform," *Harvard Educational Review* 57 (February 1987): 6

At times, *research* was invoked, seemingly, to suggest that the FPMS was beyond reproach. When research is held up as authority that cannot be challenged, it is an example of *scientism* Scientism is "science improperly conceived, understood, and practiced."[40] In scientism, there is an attempt to extend the authority of science beyond its bounds. "Sometimes," says Sergiovanni, "such attempts are fradulent ... [when a] workshop provider ... deliberately (albeit vaguely or selectively) appeals to the 'research' to lend credence and marketability to her or his ideas and prescriptions."[41] Scientism is more prevalent in disciplines that are young and less secure in their footing with other disciplines or professions.[42] Education is thus a likely candidate.

Because of cases like *Sweeney,* scientism is finding its way into the courts and becoming sanctioned by law, and so legalized. When courts uphold the use of instruments like the FPMS, then they appear to be both legally right and educationally sound.[43]

The developers also claimed that the FPMS is generic; it can be used with all teachers in all grade levels and subject areas, despite research that shows otherwise.[44] This claim was further substantiated with norming studies done on more than 1,200 beginning teachers and 20,000 master teachers[45] The developers explained their notion of generic:

> Whether the teacher smiles one way or another, as long as it is *genuine,* makes no difference. Or again, the teacher's particular way of asking questions, as long as they are *clear,* is of no particular significance [italics added].[46]

An assumption at the heart of generic is that such teacher behaviors—selected and defined according to the developers' values—give the illusion of being

---

[40]Thomas J Sergiovanni, "Science and Scientism in Supervision," *Journal of Curriculum and Supervision* 4 (Winter 1989): 93

[41]Ibid, p 95

[42]Barry Barnes, *About Science* (Oxford. Basil Blackwell Limited, 1985), p. 99.

[43]See, for example, Helen M. Hazi and Noreen B. Garman, "Legalizing Scientism Through Teacher Evaluation," *Journal of Personnel Evaluation in Education* 2 (December 1988) 7–18 There is a parallel with the use of the concept of *job-relatedness* in the courts of the early 1980s. When states required teachers to take competency tests and teachers challenged them, the use of tests was defended based on the concept of *job-relatedness* If a test was proved job-related, then it withstood legal challenge. When a test was job-related, it became legally defensible. If claims about "the research" go unchallenged in the courts, then the concept *based on research* will also come to mean that an instrument, method, or idea is legally defensible, legitimate in the eyes of the court, and thus legitimate for educational practice.

[44]Susan S. Stodolsky, "Teacher Evaluation. The Limits of Looking," *Educational Researcher* 13 (November 1984): 11–18

[45]For example, Florida Coalition for the Development of a Performance Measurement System, *Teacher Evaluation Study. Norming Study, Generic Competence Revision, Feedback Proposal, Cognitive Examinations, and Specialized Domains (1983–84)* (Tallahassee Florida Coalition, 1984).

[46]B O. Smith, Donovan Peterson, and Theodore Micceri, "Evaluation and Professional Improvement Aspects of the Florida Performance Measurement System," *Educational Leadership* 44 (April 1987): 19.

"so basic" and therefore so simple to observe Ironically, what the developers fail to see is that a *genuine* smile and *clear* questions still involve judgment.[47]

Furthermore, some scholars believe that teaching is not generic. The work of Shulman and his colleagues involved in developing prototypes for a new generation of teacher assessments is one example.[48] Shulman "strongly disputes" the position that teaching skills are generic across ages and school subjects and that they are easily observable in classrooms settings. He argues that teaching is more than a behavioral process, that it involves both intellectual reasoning and acting, that "teaching typically occurs with reference to specific bodies of content or specific skills, and that modes of teaching are distinctly different for different subject areas.[49]

The developers further claimed that the FPMS is *objective* According to the developers, objectivity is a mathematical state that involves more than one observer—not a stance that the observer takes toward the teacher·

> Another term for what is called intercoder agreement . . . when two or more observers observe a teacher at the same time, that the scores that they derive from those observations must be substantially the same.[50]

The intent of this objectivity became clear when the developers discussed why the FPMS was devised. They compared the FPMS to annual evaluations done by principals:

> Our initial motive for entering into the development of the FPMS was, number one, the primitive state in which we were at that point in time in evaluating teachers and casting high-inference judgments on teachers that made a substantial difference in their lives. They were being either promoted or retained for tenure or dismissed on the basis of very high-inference judgments that were frequently biased by personal opinions of the teachers' value or nonvalue.[51]

---

[47]Although it is beyond the scope of this article to take issue with the content of the FPMS, the reader should note that certain values are built in to the FPMS For example, the coding of items is based on the intervening-behavior rule. When the task changes, the observer makes another tally. Thus, change in task—not the amount of time spent on any task—is valued. The teacher is thus rewarded for doing different, rather than the same, tasks. Another example concerns "normal use." Observers are not supposed to code normal use of chalkboards or overhead projectors. One of the trainers provided an example: "If you're just going to write on the chalkboard or write on the overhead as a way of displaying information, that's normal use", see *Proceedings Volume III, Sweeney v. Turlington and the State Board of Education,* Case No 86-0023, Clearwater, FL (1986, April 15), p. 84. For a more thorough critique of the FPMS, see C. J. B. Macmillan and Shirley Pendlebury, "The Florida Performance Measurement System: A Consideration," *Teachers College Record* 87 (Fall 1985): 68–78

[48]Besides Shulman, see also Walter Doyle, "Paradigms for Research on Teacher Effectiveness," in *Review of Research in Education,* vol 5, ed L S Shulman (Itasca, IL F E Peacock, 1978); F J McDonald and P. Elias, *Executive Summary Report. Beginning Teacher Evaluation Study, Phase II* (Princeton, NJ: Educational Testing Service, 1976).

[49]Lee S. Shulman, "Assessment for Teaching An Initiative for the Profession," *Phi Delta Kappan* 69 (September 1987): 41.

[50]*Proceedings Volume II, Sweeney v. Turlington and the State Board of Education,* Case No 86-0023, Clearwater, FL (1986, April 15), p. 22

[51]Ibid., p 14.

Thus, the intent of objectivity is fairness—to eliminate the high-inference judgments that appear to rule annual evaluations.

Fairness, however, is only an illusion, since "in any evaluation one may learn as much, if not more, about the evaluator than the intended subjects and objects of the evaluation."[52] McCutcheon further describes what an observation instrument reveals:

An observation instrument provides the observational framework, the perceptual lenses through which the classroom is to be viewed. . . . It may indeed reveal at least as much about the observation instrument—its assumptions, its categories, its time frame, its counting method, and so forth—as it does about the classroom. We can only see, only count, only code those items admitted to perception by the instrument.[53]

What does the FPMS reveal about the developers? First, we learn what they believe about this body of knowledge. A simplicity (and even commonsense appeal) seems evident in its maxims: "The indicators are simply elements of effective teaching, and teachers can build them into their classroom performance at will."[54] Why is this body of knowledge important? If teachers use this body of knowledge, then they can be called "professional."

Improvement of pedagogical practice requires a body of objective knowledge about teaching procedures and their effects. Only if teachers have such tested knowledge can they make proper diagnoses and select procedures that satisfy professional expectations and requirements [55]

And

Moreover, the System's research base establishes its professionalism When teachers understand the research underlying their performance and realize that what they are doing is not based on opinion or mere personal experience, they feel much more secure in their behavior and are likely to act with more enthusiasm and confidence than if what they do has no research support.[56]

Therefore, how is the FPMS used?

In the evaluation of teachers, we attempt to determine the degree to which teachers have attained such knowledge and the degree to which they apply it in their instruction.[57]

The FPMS is also said to be *precise:*

The information that's obtained is much more precise. It's not based on the judgment of one rater. It's based on relatively objective data that's collected from a large number

[52]Thomas J. Sergiovanni, "Expanding Conceptions of Inquiry and Practice in Supervision and Evaluation," *Educational Evaluation and Policy Analysis* 6 (Winter 1984): 360.

[53]Gail McCutcheon, "On the Interpretation of Classroom Observations," *Educational Researcher* 10 (May 1981): 9.

[54]B. O. Smith, Donovan Peterson, and Theodore Micceri, "Evaluation and Professional Improvement Aspects of the Florida Performance Measurement System," *Educational Leadership* 44 (April 1987): 19.

[55]Ibid., p. 16.

[56]Ibid., p. 19

[57]Ibid , p 16.

of teachers, so that the teacher being examined can be put in the rank order of the teachers being considered, rather than relying on the impressions of some one rater [58]

This quotation reveals the reason for this precision—to be able to discriminate so that only "a few" teachers can be identified as superior and so qualify for limited merit bonuses.

Shulman sums up the danger of such political expediency:

The great danger occurs, ... when a general teaching principle is distorted into prescription, when maxim becomes mandate. Those states that have taken working principles of teaching, based solely on empirical studies of generic teaching effectiveness, and have rendered them as hard, independent criteria for judging a teacher's worth, are engaged in a political process likely to injure the teaching profession rather than improve it.[59]

*False Confidence in Observing the Teaching Act*

While reading the testimony of Maryanne's observers, I was struck by comments regarding how they felt about their ability to conduct observations. They expressed a "confidence." This felt confidence took on meaning when I remembered that, in their judgment, Maryanne had taught an effective lesson.

The assistant principal felt confident in marking the instrument from training:

I felt confident to mark what was the positive and negative indicators when they were indicated in the class. I felt like I was trained to do that, I felt comfortable doing that at that time I did it. Whether I did it correctly or not, that's another thing, but I felt comfortable when I did it.[60]

The principal also expressed confidence in his training.

I felt very confident going into the appraisal session. Today I'm not so sure.[61]

The attorney for the defense called the training "very impressive." Administrators participated in three days of training. The first day was spent reviewing research on teacher effectiveness, which formed the "foundation" for the FPMS items. During the next two days, principals watched videotapes and participated in simulations.

The training resulted in what was considered rigorous testing. Principals had to take multiple-choice tests on the content of the research base and the coding manual and complete the FPMS while watching two teaching tapes. If they did not meet the criterion score, they had to be retrained.

---

[58]*Proceedings Volume III, Sweeney v Turlington and the State Board of Education,* Case No. 86-0023, Clearwater, FL (1986, April 15), p 127.

[59]Lee S Shulman, "Knowledge and Teaching: Foundations of the New Reform," *Harvard Educational Review* 57 (February 1987). 11.

[60]*Proceedings Volume I, Sweeney v Turlington and the State Board of Education,* Case No 86-0023, Clearwater, FL (1986, April 15), p 107

[61]Ibid , p. 132.

Certified trainers conducted training. Trainers went through the same three days of training and a practicum. During the practicum, they provided three days of training back in their own school districts. They were then evaluated on the pass rates of their trainees. Trainers were certified when they "demonstrated that they [could] deliver the training as it [was] designed and specified in the training manuals."[62]

A *trainer maintenance* program included updates three times a year to prevent *observer drift*. "If they go long periods of time without either conducting an observation or reviewing the coding rules, they will *drift* away from being accurate [italics added]."[63] One comment seemed to finally place the principals' false confidence in perspective. Betty Fry said: "That was the best training they've ever had in teacher evaluation; in fact, it's the only training they've ever had in how to observe teachers, how to code behaviors."[64] Indeed, training in evaluation can empower principals. With such training, however, their judgment seems to have no relationship to, or consequence in, the process of identifying superior teachers.

### *The Observer's Disenfranchisement*

The significance of the case for supervision is that the observer is disenfranchised from the judgment-rendering process. The observer is disenfranchised by the narrowly defined role. "to code what the teacher does in the classroom. . . . It's not what the coder thinks of the teacher, but what the teacher does in the classroom."[65] This narrow definition is based on the assumption that observing and evaluating are and should be separate processes and that when done by the same person results in bias against the teacher:

> So when they are trained as observers, they are trained to collect data. The process of observing and evaluating are two separate processes. You don't do both at the same time. They're done separately. You collect data to make the evaluation, and then the evaluation is done based on a set of data.[66]

This disenfranchisement is also present in the principals' access to information Because the career ladder was voluntary and because the principals were only data collectors, they did not receive a score report unless the teacher provided it A computer—not the observer—determines whether improvement is possible. A computer calculates *the hypothetical third obser-*

---

[62]*Proceedings Volume III, Sweeney v Turlington and the State Board of Education,* Case No. 86-0023, Clearwater, FL (1986, April 15), p 53

[63]Ibid., p. 49.

[64]Ibid, p 55.

[65]*Proceedings Volume II, Sweeney v Turlington and the State Board of Education,* Case No. 86-0023, Clearwater, FL (1986, April 15), p. 38

[66]*Proceedings Volume III, Sweeney v Turlington and the State Board of Education,* Case No. 86-0023, Clearwater, FL (1986, April 15), pp 64–65

*vation score* based on scores from the previous evaluations. A hypothetical third score decides whether a teacher has a chance to improve the score and thus warrants a third evaluation.

When observer judgment is allowed into the process, it awards *bonus points*. Principals may award 10 points if they believe the teacher was "exceptional." The use of bonus points is acknowledged as "purely subjective." Still, principals can refer to noninstructional performance—extracurricular activities, professional organizations, and communications with parents—when awarding the points. But the points must be approved and then are mathematically worth only 10 percent of the teacher's median score.

Ironically, this disenfranchisement from the judgment-rendering process can benefit the observers, especially when they can defer judgment to a book. For example, when asked whether using an overhead projector while lecturing counted as "normal use" in the FPMS system, Betty Fry responded, "It doesn't matter what I think, it's the coding manual that governs how the item is coded."[67] Another example comes from Ann Wilson, also a trainer When asked whether she believed that the amount of time in an instructional format should affect coding, she reponded, "I don't think what I believe matters."[68]

*The Dichotomy Between Observation and Judgment*

Because the observer is disenfranchised from the judgment-rendering process, an unnecessary dichotomy exists between observation and judgment. Robert Soar, an evaluation expert brought in to testify for the state, best describes this dichotomy:

> The typical rating scale asks an observer to make an evaluation without measuring, but the observer using a structured low-inference observation schedule measures without evaluating. That distinction, I think, is the heart of the issue.[69]

At the heart of the dichotomy is the assumption that judgment computed arithmetically is error-free and thus more sound.

The state built a case for this dichotomy by pointing out the flaws of annual evaluation in comparison with the FPMS. One finding of the hearing indicates why the state's argument was successful:

> The evidence demonstrates that the ratings received on the Pinellas County School Board's annual evaluation instrument are in no way comparable to the scores received under the FPMS. The former is a method of evaluation without measurement. It does not attempt to differentiate between the quality of performance by different teachers Its purpose is *to provide feedback to teachers and to identify areas needing improvement or change* for employment purposes. The later FPMS is a system of measurement followed by comparison with norms. Only then does evaluation and identification of

[67]Ibid., p. 84
[68]Ibid., p. 99.
[69]Ibid., pp 124–125

superiority occur, which is the purpose and intent of the Master Teacher Program [italics added].[70]

Because the purpose of the annual evaluation is to provide help—"to provide feedback" and "to identify areas needing improvement"—it cannot also, apparently, identify superior teachers.

*The Teacher's Disenfranchisement*

The public—not the teacher—is the audience for the FPMS and the Master Teacher Program. The instrument and the program were designed to assauge citizens, legislators, and parents who fear that Florida children are being taught by incompetent teachers. Shulman best describes this fear as:

> teachers who do not teach, or teach only what they please to those who please them, who prefer the transient kicks of frills and fads to the tougher, less rewarding regimen of achieving tangible results in the basic skills; who close their school house doors and hide their incompetence behind union-sheltered resistance to accountability and merit increases; whose low expectations for the intellectual prowess of poor children lead them to neglect their pedagogical duties toward the very groups who need instruction most desperately, or whose limited knowledge of the sciences, mathemat ics, and language arts results in their misteaching the most able.[71]

Ironically, Maryanne Sweeney was handed a copy of the FPMS at a fall meeting before her observation, but she could not obtain any information about it—not even from her husband, a trained observer. Maryanne never even received a copy of the completed instrument until her attorney obtained it as part of the legal challenge to the FPMS. All she received was a report form indicating she scored below the 50th percentile. She then had to write a letter to request her exact score. Perhaps some believe that when teachers volunteer for such programs, they leave their rights at the classroom door

## THE AFTERMATH

The career ladder was repealed as of July 1, 1987, primarily because of opposition from teacher associations and poor teacher morale Timar and Kirp capture the extent of this conflict:

> Florida teachers threatened to file a lawsuit when the Florida education department rejected the applications of those teachers who forgot to use zip codes, failed to use Social Security numbers, or committed other similar minor infractions on their applications for a merit pay program And nearly 3,500 applications were rejected for failure to comply with the language of the rule implementing the program, which

[70] *Sweeney v Turlington and the State Board of Education*, Final Order, Case No 86-0023, Department of Education, Tallahassee, FL (1986, September 22), p. 8.

[71] Lee S Shulman, "Autonomy and Obligation: The Remote Control of Teaching," in *Handbook of Teaching and Policy*, ed. Lee S. Shulman and Gary Sykes (New York: Longman, 1983), p. 484.

required a "complete application." That is, if one of fourteen items on the application form was not completed, the application was rejected. The issue was resolved when the state board of education, faced with thousands of administrative hearings and possible court appeals, instructed the department of education to give the rejected applicants another chance. State policymakers in Florida decided to avoid prolonged litigation and conflict over merit pay by abolishing their program at the end of the 1986 school year.[72]

Although the FPMS is no longer used for the career ladder, its influence should not be underestimated. It is considered a forerunner of state beginning teacher assessment programs. It has been used in Kentucky, and many of its same principles have been used in instruments to evaluate teachers in South Carolina, North Carolina, and Texas.[73] It is still being used in Florida's Beginning Teacher Program and could be resurrected for use with its inservice teachers because the state recently passed new regulations for the annual evaluation of instructional personnel.[74] In fact, in some states "locally developed school district teacher-evaluation systems are nothing more than modified versions of the state's beginning teacher assessment program."[75] Because of costs and the legal need to establish validity and reliability, "there is reason to believe that many states will only modify their existing systems developed for novices."[76]

## THE LESSONS OF THIS CASE

The case of *Sweeney* was set within the political context of educational reform. It was heard during a time of crisis in education, when student and teacher quality was suspect, when testing (in its broadest sense) became a "QIIC" fix. When the FPMS was used to accomplish the legislative mandate

[72]Thomas B Timar and David L Kirp, "Education Reform and Institutional Competence," *Harvard Educational Review* 57 (August 1987) 324 Two lawsuits were filed objecting to the Master Teacher Program on the grounds that the state interfered with the collective-bargaining process guaranteed public employees. Both cases were ruled in favor of the state. See *Florida Teaching Profession—National Education Association v Turlington*, App 1 Dist, 490 So 2nd 142 (1986); *United Teachers of Dade FEA/United, AFT, Local 1974, AFL-CIO v Dade County School Bd*, App. 1 Dist., 472 So. 2nd 1269 (1985).

[73]Linda Darling-Hammond and Barnett Berry, *The Evolution of Teacher Policy* (Washington, DC. Center for Policy Research and the Center for the Study of Teaching, Rand Corporation, 1988) See also H Tyson-Bernstein, "The Texas Teacher Appraisal System· What Does It Really Appraise?" *American Educator* 11 (Spring 1987). 26–31

[74]"Approval of District Performance Assessment Systems," *Florida Administrative Code*, Rule 6B-4 0042(1), (1988).

[75]Linda Darling-Hammond and Barnett Berry, *The Evolution of Teacher Policy* (Washington, DC. Center for Policy Research and the Center for the Study of Teaching, Rand Corporation, 1988), p. 34

[76]Ibid., p. 35

for recognizing superior teachers, it ceased being just "a reliable, valid, and normed performance-evaluation system" and became a political tool.[77]

A series of claims helped give the appearance that the use of the FPMS was beyond reproach. Claims that the FPMS was based on teacher-effectiveness research and that it was generic, objective, and precise helped the state present a convincing legal argument. These claims help to perpetuate *the myth of the objective instrument*—the existence of a tool that can accurately and precisely measure teacher performance in the classroom—as if one instrument exists that can satisfy teachers' concern for fairness and due process, administrators' concern for easy use and minimal training, and the public's concern for selecting superior teachers.

When an evaluation tool such as the FPMS withstands legal challenge, it appears to carry a stamp of approval and to be beyond legal reproach. More important, it appears "right." Some may think that if it is "right" in the eyes of the court, it becomes a prescription for the educational community. Administrators, searching for the one uniform instrument or method to evaluate all teachers, especially welcome legal sanctioning because their evaluations must comply with due process.

Using an instrument such as the FPMS gives the illusion of replacing human judgment with machine judgment, as if a computer was more fair, educationally sound, and error-free. Even the attorney for the state said that "it"—as if some neutral instrument, not a person—fairly and reliably assessed Sweeney's classroom performance.

In reality, instruments such as the FPMS do involve human judgment: first in deciding which teaching behaviors to include in and exclude from an instrument and second in deciding which in-classroom examples to use to score it. Finally, human judgment is involved in establishing a scoring system (that administrators could not understand or abuse). Human judgment has merely been sanitized by a machine.

Ironically, in subsequent interviews, Maryanne believed that the instrument was not in question, but rather the judgment of her observers was in error. Also, her observers still could not explain the discrepancy between her score on the instrument and their judgment about the effectiveness of her lesson. One blamed the scoring system for the discrepancy. In both instances, the practitioners remain mystified about their experience with the instrument, and their roles have the potential to remain adversarial, where teachers are objects of, not participants in, the evaluation process.

Who is really the culprit and heroine in this drama? Unavoidably, it is human judgment. We must now respond to Sergiovanni's call to admit that

[77]For current perspectives on the political nature of testing, see Nancy Cole, "Testing and the Crisis in Education," *Educational Measurement: Issues and Practices* 3 (Fall 1985): 4–8; E. B. Fiske, "America's Testing Mania," *Education Life, Supplement, The New York Times*, April 1988, pp. 16–20; Blake Rodman, "Testing Practicing Teachers: The Battle Nobody Really Won," *Education Week*, 16 March 1988, pp. 1, 13.

subjectivity does exist in evaluation."[78] Our alternative is to live with systems such as the FPMS that diminish the evaluator's role; that reduce, then measure, observable yet trivial behaviors; and that result in distorted pictures of teaching.

Instead, our goals should be to reunite, not divide, data collection and judgment and to evaluate teaching. Evaluation is "a rational process, involving human discernment and judgments within a value system and a context."[79] Like teaching, the evaluation of teaching involves both "reasoning" and "acting." In the judgment-rendering process, our reasoning can be tested with criteria such as whether it is based on sufficient evidence and defensible assumptions, whether it can be corroborated with available knowledge, and whether it has utility for the teacher as well as the supervisor.[80]

We must also replace positivistic concepts in the field with ones that acknowledge the value of human judgment. For example, we need an alternative to objectivity. Why not *receptivity?* Receptivity is an open-minded stance that the observer takes toward teaching ideas and techniques that are different from those preferred. Receptivity comes from the supervisor recognizing and accounting for the espoused platforms of both teacher and supervisor.

Finally, we must replace rationalistic conceptions of teaching with more real-to-life ones. Classrooms, lessons, students, and teachers are more complex and less generic than an instrument like the FPMS would have us believe. When we use such instruments, we can only explain lessons like Maryanne Sweeney's as an *aberrant observation*—an unusual observation at the high or low end of the scale.[81]

## AN ASIDE

At the 1986 meeting of the American Educational Research Association, Division K sponsored a symposium to compare different systems to analyze teaching. The participants included Barak Rosenshine, Donovan Peterson, Roger Shuy, Elliot Eisner, Sara Delamont, and Greta Morine-Dershimer, its organizer. They analyzed a videotape of William Bennett, former U.S. Secretary of Education, teaching the Federalist Paper No. 10 to a history class at a District of Columbia high school. When Peterson used his FPMS to evaluate Bennett, he scored at the 50th percentile and thus would not have qualified for the Master Teacher Program in Florida either.[82]

---

[78]Thomas J. Sergiovanni, "Will We Ever Have a True Profession?" *Educational Leadership,* 44 (May 1987): 44–49.

[79]Ibid., p. 48

[80]I am borrowing from the work of Gail McCutcheon, "On the Interpretation of Classroom Observations," *Educational Researcher* 10 (May 1981). 5–10

[81]*Proceedings Volume II, Sweeney v Turlington and the State Board of Education,* Case No 86-0023, Clearwater, FL (1986, April 15), p. 42

[82]Donovan Peterson, Jeffrey Kromrey, Theodore Micrri, and B O. Smith, "Florida Performance Measurement System: An Example of Its Application," *Journal of Educational Research* 80 (1986): 141–147

As the participants described Bennett, they revealed the diversity of ways of seeing in classrooms. This experience led Bennett to later conclude, "If nothing else this demonstration should caution us in our move toward teacher assessment to preserve some flexibility in applying particular assessment techniques."[83] Referring specifically to items on the FPMS that did not apply to his lesson, Bennett aptly notes:

> For purposes of feedback and guidance, an instrument such as this may be helpful, but if used as a rating device, I as a teacher would want to be sure that reasonable deviations from standard practice could be accommodated. Perhaps some portion of our assessment devices should include alternate tracks for teachers who use different routes in accomplishing the same goals.[84]

---

HELEN M. HAZI is Associate Professor, Education Administration, West Virginia University, 606 Allen Hall, Morgantown, WV 26506-6122.

Oliva, Peter F. *Supervision for Today's Schools*, 3rd ed. New York: Longman, 1989. 602 pp. $30.95.

This synoptic text has sections on helping teachers improve instruction, develop curriculum, and learn through staff development. Supervision issues, roles, and evaluation are also presented along with extensive references to supervision literature, numerous activities for further study, and explicit objectives for each chapter

Sinclair, Robert L., and Sonia M. Nieto, eds. *Renewing School Curriculum: Concerns for Equal and Quality Education*. Amherst, MA: Coalition for School Improvement, 1988. 93 pp.

Curriculum renewal is the subject of this book from a schools-university partnership in western Massachusetts. Essential elements are inclusiveness and local initiatives in creating productive learning environments. Promising programs and policies are presented.

---

[83]William Bennett, "Secretary Bennett's Response to Evaluations of His Teaching," *Teaching and Teacher Education* 2 (1986): 333.

[84]Ibid.