In 1996, when Dwyer and Stufflebeam last surveyed teacher evaluation practices and philosophies, they made the following assertion:

Teacher evaluation is a highly controversial area, with myriad stakeholders and a wealth of technical, psychological, political, ethical, and educational complexities. Teacher evaluation is relevant to every segment of the educational system, and society at large has an intense interest in how it is carried out and what its impact on education and on individuals’ lives will be. Thus, the criticisms of theory and practice are strongly held, and how (or whether) these criticisms are resolved has direct implications for the quality of American schooling. (Dwyer & Stufflebeam, 1996, p. 768)

They then reported on a broad range of issues and challenges raised by the nexus of the burgeoning reform movement in education, the dynamic and complicated nature of teaching as a profession, and the technical standards for high quality personnel evaluation of any kind. As their introductory matrix of “Types of Evaluations and Decisions Involved in Preparation, Licensing, Employment, and Professionalization of Teachers” makes clear (see Appendix A), there was little consensus about what practices worked best in ensuring the overall quality of the education workforce. In addition, that teacher evaluation practices were deficient and inadequate was presented as a fact; the remainder of their survey in one way or another dealt with various aspects of improvement in these practices.

Years have passed, and nothing has occurred that would undermine the assertions in the quotation above. There are, however, two significant areas of change, in the public policy arena and in assessment methodologies, that could have profound effects on teacher evaluation practices in the next decade. The single most important shift in the public policy arena since Dwyer and Stufflebeam’s survey is the emergence of a tidal wave of support for what is
loosely called “teacher accountability.” What this seems to mean in effect is a growing insistence on measurement of teacher quality and teacher performance in terms of student achievement, which is all too often poorly defined, crudely measured, and unconnected to what educators regard as significant learning. While there have been numerous explorations of varied approaches to teacher evaluation that could lead to increased sophistication, validity, or utility of teacher evaluation practices in schools by school personnel over the past decade, none of these has resulted in widespread application of new understandings of evaluation practice for teachers.

Because there is still little consensus about acceptable ways to meet the very substantial challenges posed by links between measures of student achievement and consequent conclusions about teacher effectiveness, the fact that this issue dominates current discourse about teacher evaluation is very significant, and somewhat alarming. This is not a new effort or a new issue, but the heated insistence on its power as the single most important criterion for establishing a teacher's effectiveness is new (Millman, 1997, pp. 6–7). Simply put, most efforts to connect student achievement to individual teacher performance have foundered in the past on these weaknesses:

- The measurement does not take into account teaching context as a performance variable
- The measurement is unreliable, in part because it does not include time as a variable – both the teacher’s time with a cohort of students; and some model or models of sufficient time to see learning effects in students.
- The measures used to reflect student achievement are not congruent with best practice and philosophy of instruction in modern education.

Several notable efforts to address these deficiencies have been launched in the past decade; we discuss those below.

The second area of change has been in the area of teacher assessment. Here there is both good news and bad news. Teacher testing is now more than ever a high stakes enterprise. The most striking example of the potential stringency of teacher testing practices is provided by the experience of teachers in Massachusetts, which introduced a new teacher licensing testing program in 1998, the Massachusetts Educator Certification Tests (MECT). At its first administration in April 1998, only 41 percent of all test takers met the passing standard established by the state on all three parts of the test: reading literacy, writing literacy, and subject matter knowledge (Haney et al., 1999). There have been and continue to be numerous analyses of the technical and policy implications of the development, implementation, and continuing use of the MECT (Haney, et al., 1999; Ludlow, 2001; Melnick & Pullin, 1999; Melnick & Pullin, 2000; Luna, Solsken, & Kutz, 2000), but the essential point of the very public suffering of Massachusetts educators is how congruent this testing initiative and the responses to its results are with the rhetoric of “teacher accountability.” All of the furor over the Massachusetts approach to teacher testing emphasizes in one