mistakes waiting to be made—any one of which, all by itself, could scuttle a state's otherwise wonderful teacher-evaluation system.

So that you can be on watch for these four implementation mistakes, let's explore each of them briefly before wrapping up this chapter. Although each of the four mistakes, if made, can cripple a teacher-evaluation system, none of them needs to be made. And, happily, all of these mistakes, if they have already been made, can be corrected.

Mistake 1: Using Inappropriate Evidence of a Teacher's Quality

Poorly chosen evidence. A state's educational officials, those who attempt to plan and implement a federally spawned teacher-evaluation system, rarely specialize in designing or utilizing the evidence-collection tools needed to make evaluative decisions about teachers. Given the federal advocacy of multiple measures for determining a teacher's effectiveness, it will surely be necessary to employ diverse assessment techniques when collecting quality-illuminating evidence, that is, the data or documentation that helps teacher evaluators get an accurate fix on a given teacher's skill. Among the sorts of evidence-collection instruments that state officials will usually find they'll need are (1) observation instruments to view a teacher's classroom activities, (2) rating forms to collect evidence from the administrators or students who see a teacher in action, and (3) achievement tests such as statewide standardized tests or teacher-made classroom assessments so that we can measure students' knowledge and skills.

Yet, even though a number of state departments of education may have on their staffs a number of assessment specialists (increasingly, these days, this is a small number), in most instances those staff members are not familiar with constructing and using data-gathering tools intended to *evaluate* teachers. When federal officials call for multiple evidence sources, they are clearly asking for creation of multiple sorts of appropriate evidence, not more variations of misleading evidence. But almost all state departments of education, regrettably, possess few experienced teacher evaluators familiar with the sorts of evidence-based evaluative systems called for by the recent federal teacher-evaluation initiatives.

To illustrate, one of the sources of evaluative evidence often employed in the evaluation of teachers is information collected

via classroom observations. But carrying out accurate observations regarding what goes on in a teacher's classroom is much more complicated than is usually thought. For example, many state teacher-evaluation systems now incorporate some form of classroom observations in their programs, and have based their approaches on one or more of the widely used classroom-observation systems employed in various parts of the nation. (Two of the most popular of these observation systems will be described in Chapter 5.) But state-decreed modifications in those already established observation protocols can sometimes cause serious damage to the accuracy of the evidence they yield.

Implementation Mistake 1 is, perhaps, the most insidious of the four implementation mistakes we'll be considering. This is because, in many instances, the wrong evidence-collection instruments will actually *look like* the right evidence-collection instruments. As a brief example of this first implementation mistake, let's consider a misstep that's apt to be made in many states, especially because of federal insistence on using evidence of students' growth as a "significant factor" when appraising teachers.

Due to the ready availability of students' scores on a state's annual NCLB-mandated accountability tests, many states will be inclined to use students' year-to-year performances on those accountability tests in order to calculate some sort of "growth indicator" for all teachers whose students complete each year's NCLB tests. However, as will be described in more detail in Chapter 4, in almost all instances there is currently *zero* evidence at hand indicating that these state accountability tests yield data permitting valid inferences about a teacher's instructional quality. That's right; there is no evidence that those tests provide accurate indications of how well a teacher can teach!

Can you see that if the tests being used to measure students' growth do not permit us to make valid inferences about a teacher's instructional effectiveness, then it is absurd to evaluate teachers by using the results of such tests, especially when those results are to be regarded as a significant evaluative factor? Because too many designers of state-level teacher evaluations inaccurately believe that "a test is a test is a test," we are likely to see many states heading up their teacher-evaluation collection of evidence by using students' scores on the wrong kinds of achievement tests. Tests that, based on the performances of a teacher's students, can't demonstrably help us

identify a teacher's competence should function as an insignificant rather than significant evaluative factor.

Summing up, then, mistakenly employing incorrect evidence to evaluate teachers is the first, and perhaps most important, implementation mistake likely to be made as states undertake their new teacher-evaluation programs. Because of the convenient availability of students' test scores on states' NCLB accountability tests, many state officials will be inclined to employ such evidence prominently even though the use of many accountability-test scores for a teacher-evaluation function may be unsupported.

Mistake 2: Improperly Weighting Evidence of a Teacher's Quality

So, if Implementation Mistake 1 is, in essence, "using the wrong evidence," what about the weighting of that evidence? Weighting mistakes take place when inappropriate evaluative significance is assigned to different sources of evidence such as (1) students' test performances, (2) administrator ratings of teachers' skills, (3) classroom observations, and (4) parental ratings of their children's teacher. Typically, these weightings of the evaluative importance of various kinds of evidence to be used in teacher evaluation are made at the state level by state authorities—usually in consultation with concerned constituencies such as teachers' unions, parent groups, and so on. Even so, of course, weighting mistakes will be made.

A weighting mistake occurs when a given source of evidence is given either far greater, or far lesser, evaluative importance than it should be given. For example, let's assume that a state's education officials have chosen to use only three sources of teacher-evaluation evidence, namely, students' test scores, administrator ratings, and parental ratings, but have prescribed that the following weights must be used when any teacher in the state is to be evaluated:

- Students' Test Scores = 20%
- Administrator Ratings = 20%
- Parental Ratings = 60%

I am, of course, completely in favor of parents. If there were no parents, we would need no schools. However, I think that determining more than half of a teacher's evaluation by using parental

ratings of that teacher is silly. There are no predetermined "appropriate" weights to be assigned to different sources of evidence, of course, but because student-growth data must be given significance according to federal preferences, it seems obvious that evidence of students' growth on suitable tests should be assigned great weight when state architects of a teacher-evaluation framework start doing their framing. In the above illustration, a mere 20% hardly seems "significant."

No officially sanctioned guidelines exist to aid a state's officials as they wrestle with the problem of how much weight to assign to different sorts of evidence. In instances when a state's legislature has already made those weighting decisions, of course, there's no need for further state-stipulated weighting. But in settings where per-evidence weighting decisions must still be made, there are few substitutes for asking concerned constituencies to engage in open deliberations about the pros and cons of assigning weights to different kinds of evidence. Those in charge of these sorts of weighting decisions simply must do the most thoughtful, circumspect job they can in nailing down appropriate evaluative weights.

Let's turn, therefore, to another way to botch up the installation of a teacher-evaluation system—Implementation Mistake 3. This third mistake hinges on the way we should employ our quality-illuminating evidence about teachers once we've collected it. This mistake deals with the possible need to adjust the evaluative weight of the evidence to the particulars of the setting in which a given teacher is working.

Mistake 3: Failing to Adjust the Evaluative Weights of Evidence for a Particular Teacher's Instructional Setting

Particular teachers teach particular students who have been previously taught by particular colleagues in a particular school headed by a particular principal and abetted by particular levels of administrative and parental support. To evaluate different teachers as though they were operating in identical instructional settings is naïve. Yet, it is possible that some state-designed teacher evaluations will be fashioned in such a way that the cookie-cutter categories of evidence bearing on a specific teacher's caliber *must* be given equal weight regardless of the particular setting in which a teacher functions. Failure to take a teacher's instructional setting into consideration

when considering quality-illuminating evidence for that teacher, then, constitutes Implementation Mistake 3.

In the following chapter, we'll dig into the issue of making adjustments in evidence-interpretation based on differences in teachers' instructional settings. It would be delightful, of course, if all sorts of teacher-quality evidence could actually be treated in the same way—irrespective of the distinctive instructional setting in which a teacher functions. This would make an evaluator's job far easier. Nonetheless, despite the evaluative ease yielded by a never-need-to-adjust approach to evidence interpretation, by not engaging in at least some thinking about the need for teacher-specific weighting of evidence, we almost always reduce the accuracy of the evaluation we make about a particular teacher.

As an example of the need to make adjustments in the evaluative significance we give certain sorts of evidence, let's suppose that George Jarvis, because of diminishing enrollments in the elementary school at which he formerly taught fifth graders, has been transferred to another elementary school where, once more, he teaches fifth graders. To his delight, he discovers that in his new school the level of parental support for education is amazingly strong—and generally positive. In his former school, with its dwindling enrollments, most parents of the school's students were either disinterested or disgruntled, and they rarely took part in any parental initiatives to bolster what was going on in the school. In George's new school, however, annual parental engagement in various school-support activities hovers near 90% at each grade level. Now, let's suppose that one category of evidence to be mandatorily used in the evaluation of the state's teachers is parental ratings, that is, parents' responses to an anonymously completed rating form to be filled out regarding the instructional quality of their child's teacher. When this rating form was completed by parents last year for George, his ratings were middling or below. This year, in his new school, the anonymous parental ratings for George are fabulous. Clearly, when attaching significance to these parental ratings of a teacher's skill, attention should be given to the nature of the instructional setting in which the particular rated teacher operates. To treat the two sets of parental ratings as though they had been supplied by the same sorts of parents—parents who have the same view of schooling—would be foolish.

As enticing as it might be to assume that all sources of quality-illuminating evidence should be given identical weight across all

the diverse settings in which a state's teachers function, to do so would be short-sighted. This, then, is the third implementation mistake likely to ruin a state-structured teacher-evaluation program. If evidence-weighting adjustments are warranted in a particular teacher's situation, such adjustments need to be made. Let's look now at the final implementation mistake.

Mistake 4: Confusing the Roles of Formative and Summative Teacher Evaluation

More than 25 years ago, I wrote an article entitled "The Dysfunctional Marriage of Formative and Summative Teacher Evaluation" (Popham, 1988). As you can probably infer from the title, back then I thought that it was a dumb idea to mix formative and summative teacher evaluation. I still do.

Let's get these two labels properly defined. *Formative teacher evaluation* describes evaluation activities directed toward the improvement of the teacher's ongoing instruction. Formative teacher evaluation is focused on helping teachers become as instructionally effective as they can possibly be. In contrast, *summative teacher evaluation* refers to the appraisal of a teacher in a way that is aimed at making a decision about (1) whether to reward the teacher for atypically fine performance, (2) the teacher's continued employment, or (3) the need to place the teacher on an improve-or-else professional-support program. The distinction between formative and summative evaluation we owe to Michael Scriven, the philosopher and educational evaluator who, as a consequence of the enactment of 1965's ESEA, helped educators draw significant distinctions among several functions of educational evaluation (Scriven, 1967).

In a very meaningful sense, the ultimate aim of both formative and summative teacher evaluation is identical, that is, to provide children with a better education. *Formatively,* we want to improve teachers' instructional prowess so that they can do their most effective job in helping students learn. *Summatively,* we want to identify the exceptional teachers who should be rewarded as well as those teachers who, if they cannot be helped, should be relieved of their teaching responsibilities. (If you prefer, such nonimprovable teachers can be "deselected," a current euphemism for "firing" a teacher.) In all instances, either positive or negative, our actions should be taken to help children learn better. But lauding the two important,

distinct missions of formative and summative teacher evaluation is not the same thing as saying they can be conducted at the same time by the same teacher evaluator. They cannot.

I once heard a colleague, Henry Brickell, comment that a truly skilled educational evaluator was a person "who could bite the hand of the person being evaluated while appearing to be only licking it." Yes, teacher evaluators definitely need to possess more than a little hand-licking suave if they are going to be successful. But a teacher evaluator simply swimming in suave cannot *simultaneously* be a summative and a formative evaluator. That's because a teacher who needs to improve must honestly identify those personal deficit areas that need improvement. Weaknesses can't be remedied until they've been identified, and who knows better what teachers' shortcomings are than those teachers themselves? But *if* the teacher is interacting with an evaluator whose mission, even partially, may be to excise that teacher from the teacher's job, do you really think most teachers are going to candidly identify their own perceived shortcomings for such an evaluator? Not a chance!

Accordingly, although federal guidelines make it clear that support should be provided to teachers who need such support, and this means that formative teacher evaluation must be prominently employed as part of a state's overall teacher-evaluation system, this does not indicate that formative and summative teacher evaluation should necessarily be carried out at the same time by using the same evidence-collection procedures or the same teacher evaluators. Individuals who truly believe that a combined formative and summative teacher-evaluation effort can succeed are most likely to have recently arrived from outer space. They simply don't understand human nature.

It is understandable how LEA administrators and teachers might wish to inject into their summatively oriented teacher-evaluation programs serious dollops of formative teacher evaluation. It would be so much less threatening! However, as heart-warming and wonderful as improvement-focused formative teacher evaluation can appear, care must be taken so that its presence does not diminish the accuracy of summative teacher evaluation. Similarly, we dare not let the presence of summative teacher evaluation diminish the potency of formative teacher evaluation—a process from which most teachers and their students benefit substantially. The only way to avoid such contamination is to keep the two enterprises separate.

Remember, most states agreed to a serious quid pro quo when they accepted federal insistence on teacher evaluation that, in its summative aspects, was tough rather than tolerant. When federal officials signify that a state's teacher-evaluation process "will be used to inform personnel decisions," this clearly means that such evaluative systems should definitely have a *summative* function. Lacing a summative teacher-appraisal system with so much formative evaluation that it erodes the system's summative mission, then, constitutes our fourth and final implementation mistake.

The statewide evaluation of teachers has the potential to dramatically enhance or diminish the caliber of a state's public schooling. How teachers are evaluated will most certainly alter how teachers teach. Teachers, as is true with all of us, want to be regarded positively. Who among us relishes an adverse evaluation? Indeed, most teachers will make a serious effort to do what's needed in their classrooms in order to secure a positive appraisal. Consequently, such alterations in teachers' teaching will, just as certainly, influence how well or how poorly a state's students learn.

A teacher-appraisal system that inclines teachers to make good instructional decisions is likely to do just that. Conversely, a state teacher-appraisal system that points teachers in unsound instructional directions will, unfortunately, also do just that. It is for this reason that all of us need to understand enough about what's pivotal in teacher evaluation so that we can spot anything that's deficient, and then set out to improve it.

Chapter Implications for Three Audiences

As explained in the preface, each chapter will be wrapped up with three brief, paragraph-long attempts to isolate implications of the chapter's contents for each of the book's likely audiences, namely, (1) educational *policymakers* such as school-board members, legislators, and everyday citizens, especially parents of school-age children; (2) educational *administrators* such as a school district's central-office staff and school-site administrators such as principals and assistant principals; and (3) *teachers,* the individuals who are the focus of today's teacher-appraisal systems. It was pointed out in the preface that, as can be seen, a given chapter's implications for these three audiences will often be more overlapping than distinctive.