Chapter 2

Human Judgment

Needed or Not?

How important is human judgment in the appraisal of teachers? That is, when evaluating teachers, is human judgment pivotal, peripheral, or really not needed at all? That's precisely the issue we'll be considering in this chapter because, as you will soon see, when anyone sets out to design a sound teacher-evaluation process, the question of whether to incorporate human judgment in the process will invariably present itself. Indeed, the nature of any teacher-evaluation system will depend heavily on *whether* human judgment is involved and, if so, *how* it is used.

Before addressing this crucial question of whether to include human judgment as we tackle the building of a teacher-education system, however, it is important for you to confront a discouraging reality. Distressingly, no matter how much thought, planning, and even prayer we might devote to devising a foolproof teacher-appraisal program, when appraising teachers' abilities, we will always make mistakes. Teacher evaluators who yearn for an error-free teacher appraisal are destined to be disappointed. Mistakes will, unfortunately, be made.

In the previous chapter, it was pointed out that the different sources of evidence we employ when appraising teachers will vary in their evaluative significance, and so should be individually weighted according to their evaluative persuasiveness. Then, too, there are the particulars associated with a given teacher's instructional setting that may oblige us to decide whether to adjust those previously assigned weights.

In any given instructional setting, it will always be the case that meaningful variations will be present in (1) a teacher's unique personality—stemming from that teacher's idiosyncratic history of life events, (2) the particular students being taught, (3) levels of administrative support and leadership supplied to a teacher, (4) parental support of the teacher's instructional efforts, and (5) the quality of instructional textbooks and other materials. This collection of quality-relevant variables, of course, does not exhaust the ways in which teachers' instructional situations might vary. But, hopefully, you can see how each of these factors—all by itself—might distort teacher-to-teacher comparisons. Distressingly, such potentially confounding variables do not queue up in a single-file, easily isolatable fashion, each begging to be controlled or eliminated. On the contrary, these sorts of confounding factors are tightly tangled in different ways for different teachers. And is it these profoundly particularistic instructional settings that we dare not dodge when setting out to appraise any individual teacher.

By the time we try to sort out and, perhaps, compensate for the significant differences in particular teachers' instructional settings, differences that are more likely to represent a dozen such differences than merely one or two, the likelihood of accurate comparisons among different teachers becomes more and more difficult. A teacher's instructional setting matters—enormously.

As we watch SEA and LEA educational authorities in many states currently setting out to devise brand-new, more demanding teacher-appraisal systems, we need to remind ourselves that attempts to come up with flaw-free teacher evaluation are doomed to fail. And one huge reason for this impossibility flows from the varied instructional settings in which different teachers function. It is simply unrealistic to aim for flawless teacher evaluation. Accordingly, while recognizing that there will be a certain percentage of mistakes made when determining the caliber of individual teachers, what teacher evaluators need to do is make sure that the proportion of misses-to-hits is as small as it can be.

It is the myriad particulars of individual teachers that, even if recognized, make the pursuit of mistake-free teacher evaluation senseless. However, as long as teacher evaluators realize that they are engaged in what will be a sincere, but not impeccable, attempt to appraise teachers, this is a quest well worth undertaking. To the extent that teacher evaluators do as effective a job as they can, then the maximum number of students will benefit from a mistake-minimizing teacher-evaluation process. If we can be fair to teachers, and can

improve the quality of schooling they provide, then this is clearly an aspiration to be pursued.

Human Judgment's Role

If the architects of a teacher-evaluation process truly believe their teacher-evaluation procedures can be put into effect without central dependence on human judgment, then they will obviously try to devise an evaluative process that, insofar as possible, is essentially devoid of the need for judgment. Rather than trying to incorporate and *refine* any required human judgments, those who put together an evaluation system may try to dodge such judgments altogether. Designers of judgment-free evaluation systems, because their evaluative procedures will typically be as quantitative and as objective as possible, believe their approaches will be more accurate than will any evaluation strategy involving the often erroneous judgments of human beings. Such judgment-free attempts to evaluate teachers have been characterized as "people proof" (Mead, Rotherham, & Brown, 2012, p. 17). Regrettably, such judgment-free teacher-evaluation systems simply do not work. You need to understand why.

Evaluation Basics

When we evaluate someone, or when we evaluate something, our intention is to determine the quality, that is, the worth, of the person or thing being evaluated. Invariably, whatever is evaluated is being appraised regarding its goodness or badness in relation to a particular function. (Although some writers draw a distinction between "evaluation" and "appraisal," I see little difference between those two labels and will, therefore, use them interchangeably.)

When we evaluate things, such as a laptop computer, we always do so in relation to our intended use of the thing that's being appraised. For example, only an oaf would evaluate a laptop computer based on how well it keeps one's lap warm.

People, too, are evaluated in relation to particular functions. When evaluating workers, that is, when we are engaged in *personnel evaluation,* we always arrive at determinations of a person's quality by employing *evaluative criteria.* An evaluative criterion is simply a factor that's being employed to arrive at a conclusion about the quality of whoever is being evaluated. When appraising an opera singer,

for example, an evaluative criterion might be the singer's vocal range, that is, the difference between the highest and lowest notes the singer can reach. We arrive at determinations of a worker's quality by relying on one or more such evaluative criteria.

In order to see how a person stacks up against a particular evaluative criterion, we need to rely on *evidence*—evidence that's indicative of the person's evaluative-criterion status. Putting it differently, we need quality-illuminating evidence to help us decide how to appraise the individual being evaluated in relation to the particular evaluative criterion (or criteria) chosen by the personnel evaluator. It is only when we can rely on actual quality-illuminating evidence indicating how well a person satisfies a specific evaluative criterion that the evaluative criterion becomes truly useful in carrying out the appraisal of a worker's quality.

KEY PERSONNEL EVALUATION LINGO

Because of the need for clarity when evaluating teachers, the following four definitions should be understood:

- **Evaluation:** Determining the worth of a person, process, or performance.
- **An Evaluative Criterion:** A factor employed when evaluating a worker's quality in relation to a specific work-related function. A worker's quality can be evaluated using one evaluative criterion or multiple evaluative criteria.
- **Evidence:** The data or documentation chosen to give operational meaning to an evaluative criterion, that is, to illuminate a worker's quality with respect to a particular evaluative criterion.
- **Evidence Sources:** The eligible kinds of documentation or data that can be employed to ascertain an individual's quality regarding a particular evaluative criterion.

The evidence we select to indicate a person's status with respect to a particular evaluative criterion, then, spells out what the criterion actually signifies. Another way to put it is that we *operationalize* the evaluative criterion, that is, we give the criterion operational meaning by showing how we intend to represent it. Such evidence, then, illuminates an evaluated individual's quality with respect to the evaluative criterion involved.

Perhaps you can think of workers for whom a solo evaluative criterion, represented by only one source of evidence, accounts exclusively for the evaluative conclusion reached whenever such persons are appraised. In the real world of work, we find relatively few

of these solo evaluative criterion plus solo evidence-category evaluations. You will more frequently find that multiple evaluative criteria or, perhaps, multiple sources of evidence are necessary to evaluate a worker's quality. Let's look, now, at an instance in which a worker is appraised using only one evaluative criterion, but multiple evidence categories are employed to represent that evaluative criterion.

As you've seen, those who design the key features of any sort of personnel-evaluation system are the determiners of how such worker appraisals will turn out. This is primarily attributable to the decisions those designers make regarding which evaluative criteria and which evidence sources will be employed when appraising personnel. The appraisal of people in any personnel-evaluation system always boils down to the evaluative criteria chosen and the categories of evidence employed.

To illustrate how to evaluate a worker by using a single evaluative criterion, but multiple evidence sources, let's consider a professional, namely, a *physician.* When evaluating physicians, you can quickly see that one pivotal choice facing a personnel evaluator is whether to devise a general-purpose evaluation system covering physicians of every stripe, or a specialization-distinct evaluation leading to at least some difference when appraising physicians who possess different specializations, for instance, dermatologists, neurologists, or oncologists. An evaluator's choice between specialty-distinct or general-purpose personnel evaluation almost always depends on the degree to which the specializations involved are so fundamentally divergent that any attempt to evaluate all of them using a single approach would be misleading. To the extent that the dominant tasks of the specialists are essentially similar, however, a one-size-fits-all approach to personnel evaluation might well be the way to go.

To clarify this example, we can gain some insights from an acknowledged specialist in evaluating physicians—the late film star Cary Grant. In one of Cary Grant's most memorable but rarely reprised films, *People Will Talk,* Grant plays the role of a gifted physician, Dr. Praetorius, who employs a variety of decidedly unconventional techniques to improve his patients' health. In this film, when his healing techniques are questioned, Grant often declares that the *only* reason a doctor exists is "to make sick people well." And there it is, if we concur with Cary Grant's analysis, we have a solo evaluative criterion sitting there all by itself, a potent factor by which we

could evaluate all physicians. For purposes of this illustration, let's refer to this "making-sick-people-well" evaluative criterion as *wellness restoration.*

Interestingly, even though we might choose to build a physician-evaluation system that's centered solely on our wellness-restoration criterion, we could almost certainly employ different kinds of evidence when determining the degree to which a physician stacks up against such an evaluative criterion. When gauging how well a physician satisfies our wellness-restoration evaluative criterion, we could arrive at several legitimate indicators (that is, illuminators) of the degree to which a physician had satisfied the single evaluative criterion of wellness restoration.

And, of course, we might choose to vary these evidence categories according to different physicians' specialties so that the evidence sources chosen for our solo evaluative criterion (wellness restoration) meshed more appropriately with a given specialization's distinctive requirements. Clearly, the designers of personnel evaluations must make a number of important choices when setting up their evaluation strategies. This point was surely understood by Cary Grant.

What About the Evaluation of Teachers?

As teacher evaluators think seriously about the best way to evaluate teachers, what configuration of evaluative criteria and what sorts of evidence to represent those criteria will be most appropriate? This is really the point at which teacher evaluators determine the essential nature of their personnel evaluation strategy—and these strategy-shaping choices about evaluative criteria and evidence sources will surely undergird any plan to evaluate the quality of teachers.

As indicated in Chapter 1, I've been jousting with teacher evaluation for more than a half century and have often been on the losing end of those scuffles. Yet, we can always learn from losing, and I'd like to put forth my current teacher-evaluation recommendations for your consideration. I will certainly understand if you do not subscribe to my position on this issue. But, even if you disagree with me, please realize that the choices made about evaluative criteria and evidence sources will make, by far, the most difference in the way teacher evaluators try to appraise teachers. Here, then, is what I'd recommend to those who are putting together a teacher-evaluation system.

I think that the single evaluative criterion by which teachers should be evaluated is a teacher's *instructional ability.* Yes, I believe that the dominant factor to be employed in appraising a teacher should be a teacher's effectiveness in promoting worthwhile learning in students. Yet, because various kinds of relevant evidence illuminating a teacher's instructional ability can be collected, such as students' test scores or classroom observation data, this solo evaluative criterion should be operationalized via multiple evidence sources.

It should be apparent to you that, if a teacher-evaluation system were to be designed around a single evaluative criterion, but multiple evidence sources representing that criterion, then such an approach to teacher evaluation would most certainly call for substantial reliance on human judgment. Here's why.

Judgment-Requisite Choices

The following are the tasks during which teacher evaluators will need to summon their best judgment-making skills to arrive at sound conclusions about how a teacher-appraisal program ought to function:

- Selecting the evaluative criterion (or criteria) that will govern the evaluation;
- Choosing the evidence sources to illuminate each criterion chosen;
- Weighting the selected evidence sources;
- Adjusting, if needed, evidence weights according to the particulars of a teacher's instructional setting; and
- Coalescing the collection of weighted evidence.

These judgment-requiring tasks can be made by different teacher evaluators, ranging all the way from—at one extreme—a solo school principal functioning in isolation, all the way up to—at the other extreme—carefully selected review panels. Whether teacher evaluators are principals or review panels, those evaluators should all have been carefully trained for their important tasks. Ideally, teacher evaluators will also have been certificated via some sort of performance tasks for their responsibilities—but today's dearth of discretionary

financial resources for such certification-type endeavors makes this ideal rarely implementable.

The question posed at the outset of this chapter was whether, for teacher evaluation to be fair and sufficiently accurate, human judgment is really requisite. I hope you now agree with me that it most definitely is.

Chapter Implications for Three Audiences

For Policymakers: It is not uncommon for educational policymakers to take positions calling for the promotion of a particular outcome, such as the reduction of achievement gaps between majority and minority students, then assume that designated educators and their support personnel will be able to straightforwardly accomplish this policy-dictated outcome. But a key message for policymakers found in this chapter is that the evaluation of teachers is far more perplexing than is typically thought, and it must inevitably depend on the exercise of considerable human judgment on the part of those carrying out the evaluations. Policymakers who believe that mistake-free teacher evaluation is attainable are apt to be disappointed by this chapter's somber message. Hopefully, educational policymakers will recognize that, despite teacher evaluation's challenges, reliance on human judgment offers the best route to the most fair and accurate appraisal of teachers.

For Administrators: Clearly, the degree to which district and school-site administrators are able to adopt and refine the kinds of evaluative judgments called for in this chapter will depend directly on the extent to which a particular state's framework for teacher evaluation allows for district or school variations in such evaluative procedures. If a state's framework represents a "no-variations-permitted" approach, then administrators must hope that the evidence sources selected, and the evaluative weights they have been assigned, are sufficiently reasonable. Otherwise, educational administrators should seek modifications, perhaps collectively, in the way a state's teacher-evaluation procedures are supposed to operate. However, if a state framework permits district and school administrators to exercise some degree of discretion in determining a teacher's quality, then those administrators should strive to make the training of all teacher evaluators as potent as possible. The teacher-evaluation

strategy endorsed by the federal government is quite different than the teacher-evaluation approaches in place that most of today's educational administrators have experienced during their careers.

For Teachers: Because teacher-evaluation systems in which human judgment plays a prominent role is apt to permit more teacher-specific, tailored appraisals of a given teacher's quality, teachers should recognize that in most instances a judgmentally based approach to teacher evaluation will be more accurate than will judgment-free evaluation systems. Nonetheless, teachers need to recognize that all attempts to evaluate teachers, no matter how well intentioned or properly implemented, will lead to a certain proportion of mistakes. The enormous complexity of the instructional process simply makes it impossible to avoid evaluative mistakes. Accordingly, the more knowledgeable that teachers can become about the innards of the teacher-evaluation system being used to appraise them, the better positioned those teachers will be in successfully combating any inaccurate evaluations.