In Albemarle Paper Company versus Moody (1975), the company had established employee selection criteria that it found related significantly to evaluated performance on the job. When the court found that Albemarle's performance appraisal system did not focus on well-defined job tasks determined through a careful job analysis, it threw out the selection test that had been validated against performance ratings. The court ruled that a low test score was therefore not a valid basis on which to deny Moody a position he had sought. It ordered the company to compensate Moody for legal fees as well as all of the income that he would have earned if selected for the position.

In Brito versus Zia Company (1973), the employer was found to have discriminated when it laid off a disproportionate number of protected group members based on their previous performance appraisal scores. After examining the appraisal instrument, the court held that the ratings did not really reflect performance on important job tasks, but rather the "best judgments and opinions of supervisors." Such opinions were deemed an unfair basis for selective discharge, and the company was ordered to make costly amends to employees against which it had illegally discriminated.

Attributes of appraisal instruments defensible against discrimination charges are defined by the Equal Employment Opportunity Commission's "Uniform Guidelines," state departments of Fair Employment and Housing, and the courts. They suggest that performance measures be based on critical elements of the job performed and that they be both valid and reliable indicators of work. The 1978 Civil Service Reform Act goes further to require in the Federal Service that (1) employees participate in defining critical elements of their jobs, (2) employees be evaluated solely on the extent to which they fulfill these requirements—not in comparison to one another, (3) formal appraisals be conducted at least once per year, and (4) rewards be tied directly to rated performance. The model established by the act is not a bad one to follow even outside the federal government.

Validity and Reliability

Validity and reliability are two fundamental criteria for measures of anything, including work performance. *Validity* is the extent to which something really reflects what it purports to be. In terms of work performance, a valid measure is one that assesses behavior in terms of job duties or task requirements. A well-written job description, based upon careful job analysis, is the best foundation for a performance appraisal instrument. The description should clearly state major job duties in behavioral terms. For effectiveness of repair work, a valid measure

Rater biases:
- *Leniency*
- *Harshness*
- *Central tendency*
- *Halo*
- *Similarity*
- *Recency*
- *Contrast*

might be whether the machine works as it should, not whether the mechanic has a vocational school credential or an M.S. in horticulture. Valid measures of typing skill would include the number of words typed per minute and the number of mistakes made in a finished product, not pleasantness of personality or neatness of appearance.

Reliability is the consistency of a measure. If the same task is assigned the same value in repeated trials, the measurement is reliable. The bathroom scale with fatigued springs may be a valid but unreliable measure of weight. Even though it does measure pounds, it gives different readings on consecutive uses by the same person. Most unstructured interviews for employee selection are notoriously unreliable; ratings of the very same applicant have been shown to differ by rater, time of day, rater-applicant similarity, and other contextual circumstances.

Reliability problems in performance appraisals largely boil down to rater bias. Human judgment is at the heart of any system, and structuring that judgment is a major function of the appraisal instrument. Among the common rater biases are leniency, harshness, central tendency, halo, similarity, recency, and contrast errors. *Leniency* bias is the tendency of the appraiser to rate all appraisees toward the upper end of the scale, somewhere between good and outstanding, for example. *Harshness* is the tendency to rate all people lower within a range. An appraiser with *central tendency*, not surprisingly, rates everybody near the middle. All three of these amount to the appraiser using a restricted range on the scale of possible performance levels. Since appraisees can be differentiated even within a limited range, these biases pose little problem when the same rater evaluates everybody in an evaluating system; but such is seldom the case.

Problems arise from these biases when ratings from differently biased appraisers are compared. If crews A and B contain roughly similar performers, should the workers under lenient Supervisor A get larger appraisal-based raises (or more promotion opportunities) than those under central-tending Supervisor B? In any but a very small system, restriction of range tendencies, unless everybody has exactly the same one, destroys comparability across raters by allowing the same ratings or scores to take on multiple meanings. Raters do not have to be "biased" for these effects to occur. They simply might see different meanings in such commonly used evaluative words as "poor," "satisfactory," and "excellent," but the inter-rater complications are the same.

A *halo* effect is produced when an appraiser assigns the same score on each measured dimension of performance for a given employee. It often reflects an over-generalized view of the employee based on his or her true performance on a single important dimension. *Similarity* bias results in higher ratings for employees with certain attributes similar to those of the rater. This phenomenon

is very close to the one well known by its esoterically technical name, "favoritism." Appraisals based largely on work done toward the end of a work period suffer from *recency* error. Many employees make it their business to be at their best during the month preceding an annual appraisal interview. Finally, *contrast* errors derive from the tendency to rate an employee in direct comparison with another, rather than against a set of objective standards.

The structure of the appraisal instrument affects the likelihood of biases operating. Clear definitions of not only the dimensions of performance but also different levels of performance help to minimize the occurrence of errors from appraiser bias. While appraiser training can constitute another foil to these errors, considerable relevance and objectivity can be built into performance appraisal through the instrument itself.

Types of Appraisal Instruments

Many types of appraisal forms are in use. The *graphic rating scale* is by far the most widely used type. These scales come in different formats. All require the appraiser to choose the most descriptive rating or evaluative adjective from a linear graph of possibilities ranging from worst to best or vice versa. Major distinctions between formats are how the performance criteria (dimensions) and performance levels (standards or degrees) are defined.

A second type of instrument approaches performance measurement through any one of what might be termed comparative techniques. All methods of this