Chapter 6

Personnel Selection

WILL'S KASSANDRA 2.1 6 1 5

OBJECTIVES

After reading this chapter, you should be able to

1. Understand the concepts of reliability, validity, and utility.
2. Understand the validity evidence for various selection methods.
3. Discuss approaches to the more effective use for application blanks, reference checks, biographical data, testing, and various other selection methods programs in order to increase the validity and legal defensibility of each.
4. Discuss the approaches available for drug testing.
5. Describe the validity of different approaches to interviewing.
6. Explain how the various types of job candidate information should be integrated and evaluated.

OVERVIEW

It sounds simple: Match employees with jobs. Researchers have made this task easier by developing selection methods that successfully predict employee effectiveness. Still, there is a void between what research indicates and how organizations actually do personnel selection. Real-world personnel selection is replete with examples of methods that have been proven to be ineffective or inferior.

Use of validated selection models: A HPWS characteristic

Personnel selection (and retention) is key to organizational effectiveness. The most successful firms use methods that accurately predict future performance. The use of validated selection models is another of the **High-Performance Work Practices** linking this HR process to corporate financial performance. Organizations are, or should be, interested in selecting employees who not only will be effective but who will work as long as the organization needs them and, of course, will not engage in counterproductive behaviors such as violence, substance abuse, avoidable accidents, and employee theft.

A multiple-hurdle process involving an application, reference and background checks, various forms of standardized testing, and some form of interview is the typical chronology of events for selection, particularly for external hiring decisions. Internal decisions, such as promotions, are typically done with less formality. **Personnel selection** is the *process*

of gathering and assessing information about job candidates in order to make decisions about personnel. The process applies to entry-level personnel and promotions, transfers, and even job retention in the context of corporate downsizing efforts. This chapter introduces you to personnel selection, describes some of the most popular types of hiring/screening procedures, reviews the research evidence on each, and discusses the social and legal implications of the various options.

The chapter begins with an overview of measurement issues related to personnel selection and staffing. Next the various selection methods are introduced in their usual order of use. Application blanks, background checks, and reference checks are discussed first. Then the various forms of standardized tests that purport to assess applicants' suitability or KASOCs are reviewed. The use, validity, and possible adverse impact of various types of selection methods are considered, including general mental ability tests and personality tests. The final sections of the chapter discuss employment interviews and methods that have been shown to increase their validity, the use of more sophisticated (and expensive) selection procedures such as assessment centers, performance testing and work samples, and drug and medical tests in the preemployment selection process. The context of the discussion are the legal implications of the various personnel practices and pointing out where there are clear discrepancies between what typically happens in practice and what academic research indicates should happen. **This is one chapter where the distance between academic research findings and recommendations and actual selection practices is great.** The good news is that the gap is closing.

Wackenhut Security (recently acquired by G4S) had its share of selection challenges. Although recruitment efforts and a sluggish economy attracted a large number of applicants for its entry-level armed and unarmed security guard positions, there was concern about the quality of those hired and high voluntary employee turnover. The turnover rate for some positions exceeded 100 percent—meaning that the quit rate in 1 year exceeded the number of available positions. Wackenhut Security also was dissatisfied with the quality of its supervisory personnel.

The company contracted with BA&C (Behavioral Analysts and Consultants), a Florida psychological consulting firm that specializes in staffing problems and personnel selection. Wackenhut asked BA&C to develop a new personnel selection system for entry-level guards and supervisors. Underlying this request was a need for Wackenhut to improve its competitive position in this highly competitive industry by increasing sales and contracts, decreasing costs, and, most important, making certain its security personnel do the job.

The company, which already compensated its guards and supervisors more than others in the industry, wanted to avoid any increase in compensation. The company estimated that the cost of training a new armed guard was about $1,800. With several hundred guards quitting in less than a year, the company often failed to even recover training costs in sales. Wackenhut needed new selection methods that could increase the effectiveness of the guards and supervisors and identify those guard applicants who not only performed well but would be most likely to stay with the company.

First Step is Work analysis

You will recall from Chapter 4 that work analysis should identify the knowledge, abilities, skills, and other characteristics (KASOCs) or competencies that are necessary for successful performance and retention on the job. In this case, BA&C first conducted a job analysis of the various guard jobs to get better information on the KASOCs required for the work. After identifying the critical KASOCs, BA&C developed a reliable, valid, and **job-related** weighted application blank, screening test, and interview format.

The process of selection varies substantially within this industry. While Wackenhut initially used only a high school diploma as a job specification, an application blank, a background check, and an interview by someone in personnel, competitors used more complex methods to select employees. American Protective Services, for example, the company that handled security for the Atlanta Olympics, used a battery of psychological and aptitude tests along with a structured interview. Wackenhut wanted selection systems that were even more valid and useful than what their major competitors were using. Their marketing strategy would then emphasize their more sophisticated screening methods.

As with the job analysis and the recruitment process, personnel selection should be directly linked to the HR planning function and the strategic objectives of the company. For

Figure 6-1
Steps in the Development and Evaluation of a Selection Procedure

JOB ANALYSIS/HUMAN RESOURCE PLANNING
Identify knowledge, abilities, skills, and other characteristics (KASOCs) (aka: competencies).
Use a competency model tied to organizational objectives.

RECRUITMENT STRATEGY: SELECT/DEVELOP SELECTION PROCEDURES
Review options for assessing applicants on each of the KASOCs:
Standardized tests (cognitive, personality, motivational, psychomotor).
Application blanks, biographical data, background and reference checks, accomplishment record.
Performance tests, assessment centers, interviews.

DETERMINE VALIDITY FOR SELECTION METHODS
Criterion-related validation or validity generalization.
Expert judgment (content validity).

DETERMINE WEIGHTING SYSTEM FOR DATA FROM SELECTION METHODS

example, the mission of the Marriott Corporation is to be the hotel chain of choice of frequent travelers. As part of this strategy, the company developed a successful selection system to identify people who could be particularly attentive to customer demands. Wackenhut Security also had a major marketing strategy aimed at new contracts for armed security guards who would be extremely vigilant. The new selection system would be designed to identify people more likely to perform well in this capacity.

Figure 6-1 presents a chronology of our recommended strategy for selection system development and the major options available for personnel selection. The previous chapters on work analysis, planning, and recruitment have gotten us to the point of selecting job candidates based on relevant and job-related information from one or more selection methods. Each of these methods is reviewed in this chapter. But keep in mind that the focus should be on selecting or developing tools that will provide valid assessments on the critical KASOCs, competencies, and job specifications most important for strategy execution. The work analysis should identify the strategically important KASOCs or competencies from which the *job specifications* will be derived. Then particular selection methods (selection tools) should be adopted to assess people in terms of these particular job specifications.

SELECTION METHODS: ARE THEY EFFECTIVE?

This review includes a summary of the validity of each major approach to selection and an assessment of the relative cost to develop and administer each method. Three key terms related to effectiveness are **reliability, validity,** and **utility.** While these terms are strongly related to one another, the most important criterion for a selection method is *validity.* Remember the discussion of the research on **High-Performance Work Practices.** One of the HR practices shown to be related to corporate financial performance was the percentage of employees hired using "validated selection methods."[1] The essence of the term **validity** *is the extent to which scores on a selection method predict one or more important criteria.* While the most typical criterion of interest to selection and staffing specialists is job performance, companies also may be interested in other criteria such as how long an employee may stay on the job or whether the employee will steal from the organization, be violent, or be more likely to be involved in work-related accidents. But before addressing the validity of a method, let's look at one of the necessary conditions for validity: the *reliability* of measurement.

What Is Reliability?

The primary purpose of personnel selection is measuring the attributes of job candidates. A necessary condition for a selection method to be valid is that it first be **reliable. Reliability** *concerns the degree of consistency or the agreement between two sets of scores*

on some measurement device. Reliability refers to freedom from unsystematic errors of measurement. The consistency in measurement applies to the scores that derive from the selection method. These scores can come from a paper-and-pencil test, a job interview, a performance appraisal, or any other method that is used to measure characteristics and make decisions about people. The CIA uses a very long multiple-choice test as an initial screening device for job applicants to be agents. If applicants were to take the test twice 3 weeks apart, their scores on the test would stay pretty much the same (the same thing can be said for SAT scores). These tests can be considered reliable. The level of reliability can be represented by a correlation coefficient. Correlations from 0 to 1.0 show the extent of the reliability. Generally, reliable methods have reliability coefficients that are .8 or higher, indicating a high degree of consistency in scores. No selection method achieves perfect reliability, but the goal should be to reduce error in measurement as much as possible and achieve high reliability. If raters are a part of the selection method, such as job interviewers or on-the-job performance evaluators, the extent to which different raters agree also can represent the reliability (or unreliability) of the method.

Good reliability: .8 or higher

Remember our criticism about the use of graphology (or handwriting analysis) for personnel selection we discussed in Chapter 1? Handwriting analysis is used by some U.S. companies and even more European firms as a method of selection. But this method is first of all not even reliable, much less valid. If the same handwriting sample were given to two graphologists, they would not necessarily agree on the levels or scores on various employment-related attributes (e.g., drive, judgment, creativity, intelligence), supposedly measured based on a handwriting sample. Thus the method has *low reliability* as an assessment of these attributes. (But even if the two graphologists did agree on relative levels of some attribute, this agreement would not necessarily mean that their assessments are valid.)

Reliable methods tend to be long. One of the reasons the SAT, the GRE, the GMAT, and the LSAT seem to take forever to complete is so these tests will have very high levels of reliability (and they do). Reliabilities for "high stakes" tests such as the GMAT, the SAT, and the LSAT are quite high. For example, the average reliability estimates are .92, .90, and .89 for the GMAT total score, the Verbal score, and the Quantitative score, respectively.[2] But while **high reliability is a necessary condition for high validity, high reliability does not ensure that a method is valid.** The GMAT may be highly reliable, but do scores on the GMAT actually predict success in business school? This question addresses the *validity* of the method.

What Is Validity?

The objective of the Wackenhut Security consultants was to develop a reliable, *valid,* legally defensible, user-friendly, and inexpensive test that could predict both job performance and long job tenure for security guards. The extent to which the test was able to predict an important criterion such as performance was an indication of the test's **validity.** The term *validity* is close in meaning but not synonymous with the critical legal term *job relatedness*, which is discussed in Chapters 3 and 4. **Empirical** or **criterion-related validity** involves the statistical relationship between scores on some predictor or selection method (e.g., a test or an interview) and performance on some criterion measure such as on-the-job effectiveness (e.g., sales, supervisory ratings, job turnover, employee theft). At Wackenhut, a study was conducted in which scores on the new screening test were correlated with job performance and job tenure. Given a certain level of correlation, such a study would support a legal argument of job relatedness.

Validity is close in meaning to "job relatedness"

Criterion-related validity

The statistical relationship is usually reported as a **correlation coefficient.** This describes the relationship between scores on the predictor and measures of effectiveness (also called criteria). Correlations from −1 to +1 show the direction and strength of the relationship. Higher correlations indicate stronger validity. Assuming that the study was conducted properly, a significant correlation between the scores on a method and scores (or data) on some important criterion could be offered as a strong argument for the **job relatedness** of the method. Under certain circumstances, correlation coefficients even in the .20s can signify a useful method. However, higher correlations are clearly better. In general, an increase in the validity of a selection method will translate into a proportional increase in the average dollar value of the annual output from employees who are selected with this method.

While higher correlations are generally better, the size of the sample (and other factors) are very important for achieving statistical significance. Validity studies with small sample sizes will often not achieve significance mainly because of the error in the study. **Many selection methods have average validities between .20 and .40.** Samples of a minimum of 100 scores are strongly recommended in order to empirically validate in a particular setting.[3] So, do scores on the GMAT predict success in business school? Clearly, they do with an average validity of about .5 across hundreds of studies.

Another key issue that will have an impact on the results and interpretation of empirical studies is the conceptual match between a particular criterion of interest (e.g., some element of job performance) and any particular predictor. Cognitively loaded predictors (those correlated with general mental ability [GMA]) are the strongest predictors of task performance, while so-called noncognitive predictors such as personality and motivational measures are better predictors of contextual performance/citizenship behavior (e.g., effects on co-workers) and counterproductive behavior (e.g., employee theft).

A critical concept related to validity is **generalizability.** This term refers to the extent to which the validity of a selection method can generalize to other employment settings or situations. At the most basic level, generalizability concerns whether the validity of a selection method established based on a study or studies in other situations can be inferred for a new situation in which no new correlational data are collected. **Validity generalization (VG)** invokes evidence from past studies on a selection method that is then applied to a new and similar setting. Many studies have used appropriate scientific methods to establish the validity and generalizability of constructs, such as cognitive or general mental ability and emotional intelligence, and also particular instruments and methods developed to measure these constructs. **Meta-analytic** techniques are used to establish VG for a method. **Meta-analysis is a methodology for quantitatively accumulating results across studies.** Meta-analytic findings are generally more reliable than results obtained from an individual study and help researchers draw conclusions. Like other areas of scientific inquiry, meta-analytic methods have evolved and new refinements continue to emerge. These improvements have increased the accuracy of meta-analytic methods and estimates of the validity of these particular selection tests and methods.[4]

Validity Generalization

VG is an excellent alternative to empirical validation for selection methods when a criterion-related validation study cannot be done because of inadequate sample sizes or other reasons. Employers could invoke an appropriate VG study to argue that a particular test or method is valid for their setting as well. This approach is recommended if there is insufficient data to allow for an empirical study by this employer (i.e., at a minimum, less than 100 pairs of scores on an instrument correlated with performance data on the same individuals).

A VG argument for validity can be invoked if an organization can first locate previously conducted empirical studies showing that the same or similar methods (e.g., tests) are valid for a particular job or purpose. The organization should then produce an analysis showing that the job for which the method is used (or will be used) for selection is the same as, or very similar to the job(s) that were involved in the empirical studies of the VG study and that the criterion measures used in the VG studies are also important for the organization. Does an accredited MBA program need to do another study showing the validity of the GMAT for that particular program? Almost certainly not; there is plenty of evidence documenting the VG of this test for predicting business school success.

Figure 6-2 presents a summary of the meta-analytic evidence for the most popular selection tools, plus the relative cost of their development and administration. An obvious and critical question is "How large must a correlation be?" Correlations between of .20 and .30 are often discounted because they account for less than 10 percent of the variance in performance. However, as a matter of fact, a correlation of say .30 for a selection method is sufficiently large that hiring applicants who score better on this particular measure can actually double the rate of successful performance. For example, with validity at .30, 67 percent of individuals who score in the top 20 percent on a measure would have above-average performance versus only 33 percent of individuals who score in the bottom 20 percent.

Figure 6-2 Selection Tools, and Cost for Development and Administration

Tool	Validity[1]	Costs (Development/ Administration)[2]
General mental ability tests (or GMA) measure mental abilities such as reading comprehension, verbal or math skills.	.5–.7[3]	Low/low
Structured interviews measure a variety of skills and abilities using a standard set of questions.	.4–.45	High/high
Unstructured interviews measure a variety of skills using questions that vary from candidate to candidate and interviewer to interviewer.	.2–.3	Low/high
Work samples/performance tests measure job skills using the actual performance of tasks as on job.	.3–.4	High/high
Job knowledge tests measure bodies of knowledge required by a job.	.4–.5	High/low
Personality Testing[4]		
Conscientiousness	.25–.3	Low/low
Extraversion	.15–.35[5]	Low/low
Emotional Stability	.1–.3	Low/low
Agreeableness[6]	.1–.2	Low/low
Openness to Experience	.1–.2	Low/low
Biographical information measures a variety of skills and personal characteristics through questions about education, training, work experience, and interests.	.3–.4	High/low
Measures of work experience (e.g., "behavioral consistency")	.3–.4	High/low
Situational judgment tests measure a variety of skills with short scenarios (either in written or video format) asking test takers what would be their most likely response.	.3–.4	High/low
Integrity tests measure attitudes and experiences related to a person's honesty, dependability, and trustworthiness.	.3.–.4	Low/low
Assessment centers measure KASOCs through a series of work samples/exercises with trained assessors (may include GMA and other tests).	.3–.45	High/high
Reference checks provide information about an applicant's past performance or measure the accuracy of applicants' statements on their résumés.	.2–.3	Low/low

[1]Validities range from 0 to 1.0; higher numbers indicate better prediction of job performance. Ranges are reported here.
[2]References to high or low are based on relative comparisons to other methods.
[3]Validities for more complex jobs tend to be higher for GMA.
[4]Validities for personality measures tend to vary with the job. FFM self-report validity ranges reported here. Much stronger validities (.5–.6 range) for peer-based (versus self-reported) measures of personality.
[5]Stronger validity in predicting managerial and/or leadership performance; weak validities for jobs involving less interaction.
[6]Low validity for managerial jobs (.10); higher validities for team-based settings.

Sources: Adapted from W. F. Cascio, and H. Aguinis (2011). *Applied psychology in human resource management*. Upper Saddle River, NJ: Prentice Hall; and A. M. Ryan & N. T. Tippins, (2004). Attracting and Selecting: What Psychological Research Tells Us. *Human Resource Management,* 43, 307–308.

Content validity assesses the degree to which the contents of a selection method (i.e., the actual test or instrument items or components) represent (or assess) the requirements of the job. This approach to validation is of course ideal when the employer lacks an adequate sample size to be able to empirically validate a method. Subject matter experts are typically used to evaluate the compatibility of the content of the method with the actual requirements of a job (e.g., is the knowledge or skill assessed on the test compatible with the knowledge or skill required on the actual job?). Such a study or evaluation by experts also can be offered as evidence of job relatedness, but the study should follow the directions provided by the Supreme Court in *Albemarle v. Moody* (see Chapter 3) and, just to be safe, comply with the *Uniform Guidelines on Employee Selection Procedures* (UGESP). (See www.eeoc.gov for details on the UGESP.)

A knowledge-based test for "Certified Public Accountant" could be considered to have content validity for an accounting job. Many organizations now use job simulations or work samples where an applicant is instructed to play the role of a job incumbent and perform tasks judged to be directly related to the job. Content validation is ideal for these types of methods. Of course, with this approach to validation, it is assumed that job candidates have the essential KASOCs at the time of assessment. Another possible problem is

that content validation relies on the judgments of humans regarding "job relatedness" or the validity of these methods and the underlying items of the method. This approach is also inappropriate for tests of basic constructs such as cognitive or general mental ability or personality characteristics.

What Is Utility?

Low SR is needed for high utility

The validity correlation coefficient can also be used to calculate the financial value of a selection method, using a utility formula, which can convert correlations into dollar savings or profits that can be credited to a particular selection method. A method's *utility* depends on its validity but on other issues as well. For example, recall the discussion of **selection ratio** in Chapter 5. **Selection ratio is the number of positions divided by the number of applicants for those positions.** A test with perfect validity will have no utility if the selection ratio is 1.0 (one applicant per position). This is why an organization's reputation, its recruitment programs, and other HR issues such as compensation are so important for personnel selection. Valid selection methods have great utility for an organization only when that organization can be selective based on the scores on that method.

Utility (U) or expected return based on using a particular selection method is typically derived based on the formula where $U = N_s r_{xy} SD_y Z_x - N_T(C)$ where N_s = number of job applicants selected; r_{xy} = the validity coefficient for the method; SD_y = standard deviation of job performance in dollars and Z_x = average score on the selection method for hired (a measure of the quality of recruitment); N_T = number of applicants assessed with the selection method and C = cost of assessing each job candidate with the selection method. In general, the higher the validity of a method, the higher its utility. Any increase in the validity of a selection method translates into an increase in the average dollar value of the annual productivity by employees who are selected with the method. Even an increase in a small percentage can translate into a substantial annual output per employee and thus large financial gains.

Selection methods with high validity that are relatively inexpensive are the ideal in terms of utility. Before contracting with BA&C, Wackenhut Security had studied the options and was not impressed with the validity or utility evidence reported by the test publishers, particularly in the context of the $10–$15 cost per applicant. This was the main reason Wackenhut decided to develop its own selection battery.

BA&C investigated the validity of its proposed new selection systems using both criterion-related and content-validation procedures. This dual approach to validation provides stronger evidence for job relatedness and is more compatible with the **Uniform Guidelines** issued by the EEOC. The BA&C study recommended that new methods of personnel selection should be used if the company hoped to increase its sales and decrease the costly employee turnover. The resulting analysis showed substantial financial benefit to the company if it adopted the new methods for use in lieu of the old ineffective procedures. The first method that BA&C considered was the *application blank*.

APPLICATION BLANKS AND BIOGRAPHICAL DATA

Like most companies, Wackenhut first required an application blank requesting standard information about the applicant to be completed, such as his or her previous employment history, experience, and education. Often used as an initial screening method, the application blank, when properly used, can provide much more than a first cut. However, application blanks, as with any other selection procedure used for screening people, fall under the scrutiny of the courts and state regulatory agencies for possible EEO violations. HR managers should be cautious about using information on an application blank that disproportionately screens out protected class members, and they must be careful not to ask illegal questions. The **Americans with Disabilities Act (ADA)** stipulates that application blanks should not include questions about an applicant's health, disabilities, and worker's compensation history.

Application blanks obviously can yield information relevant to an employment decision. Yet, it is often the weight—or lack of weight—assigned to specific information by particular decision makers that can undermine their usefulness. Decision makers often disagree about the relative importance of information on application blanks. For instance, they might disagree about the amount of education or experience required. Wackenhut required a bachelor's degree in business or a related discipline for the supervisory job. This criterion alone, however, should not carry all the weight. Wackenhut's personnel staff made no effort to develop a uniform practice of evaluating the information on the forms. They did not take into consideration indicators such as the distance an applicant lived from the workplace. A great distance might indicate that, relative to other responses, the candidate is more likely to quit as soon as another job comes along that is closer to home.

A Discrepancy between Research and Practice: The Use of Application Blanks and Biographical Data

What companies do to evaluate application blank data and biographical information and what research suggests they should do are worlds apart. Scholarly research shows that when adequate data are available, the best way to use and interpret application blank information is to derive an objective scoring system for responses to application blank questions.[5] The system is based on a criterion-related validation study, resulting in a **weighted application blank (WAB),** with the weights derived from the results of the research. A criterion-related validation study means that the responses from the application blanks are statistically related to one or more important criteria (e.g., job tenure or turnover) such that the critical predictive relationships between WAB responses and criterion outcomes (e.g., performance, turnover) can be identified. For example, BA&C was able to show that where a security guard lived relative to his assigned duties was indeed a significant predictor of job turnover. Another useful predictor was the number of jobs held by the applicant during the past 3 years. Figure 6-3 shows some examples from a WAB. The number and sign in parentheses is the predictive weight for a response. For example, you would lose five points if you had to travel 21 or more miles to work (see #2).

The process of statistically weighting the information on an application blank enhances use of the application blank's information and improves the validity of the whole process. The WAB is simply an application blank that has a multiple-choice format and is scored—similar to a paper-and-pencil test. A WAB provides a predictive score for each job candidate and makes it possible to compare the score with that of other candidates. For example, the numbers in parentheses for the WAB examples in Figure 6-3 were derived from an

Figure 6-3
Examples of WAB and BIB

WAB EXAMPLES

1. How many jobs have you held in the last five years? (*a*) none (0); (*b*) 1 (15); (*c*) 2–3 (11); (*d*) 4–5 (23); (*e*) over 5 (25)
2. What distance must you travel from your home to work? (*a*) less than 1 mile (15); (*b*) 1–5 miles (13); (*c*) 6–10 miles (0); (*d*) 11–20 miles (23); and (*e*) 21 or more miles (25)

BIB EXAMPLES

How often have you made speeches in front of a group of adults?

How many close friends did you have in your last year of formal education? A. None that I would call "close." (20.5); B.1 or 2. (20.2); C. 3 or 4. (0); D. 5 or 6. (0.2); E. 7 or 8 (0.5); F. 9 or 10 (0.7); G. More than 10 (1.0)

How often have you set long-term goals or objectives for yourself?

How often have other students come to you for advice? How often have you had to persuade someone to do what you wanted?

How often have you felt that you were an unimportant member of a group?

How often have you felt awkward about asking for help on something?

How often do you work in "study groups" with other students?

How often have you had difficulties in maintaining your priorities?

How often have you felt "burnt out" after working hard on a task?

How often have you felt pressured to do something when you thought it was wrong?

Source: Adapted from C. J. Russell, J. Matson, S. E. Devlin, and D. Atwater, "Predictive Validity of Biodata Items Generated from Retrospective Life Experience Essays," *Journal of Applied Psychology* 75 (1990), pp. 569–580.

actual study showing that particular responses were related to job tenure (i.e., coded as either stayed with the company for over 1 year or not). Thus, applicants who had only one job in the last 5 years (#1 in Figure 6.3) were more likely to stay over a year while applicants who indicated that they had had over five jobs in the last 5 years were much less likely to remain on the job for a year or longer.

Biographical information blanks (BIBs) are similar to WABs except the items of a BIB tend to be more personal with questions about personal background and life experiences. Figure 6-3 shows examples of items from a BIB for the U.S. Navy. BIB research has shown that the method can be an effective tool in the prediction of job turnover, job choice, and job performance. In one excellent study conducted at the Naval Academy, biographical information was derived from life-history essays, reflecting life experiences that were then written in multiple-choice format (see Figure 6-3).[6] BIB scoring is usually derived from a study of how responses relate to important criteria such as job performance. Asking job candidates to elaborate on responses to BIBs with details of experiences such as dates and people involved in the events appears to enhance the effectiveness of the method by reducing the response faking (and embellishments). For example, applicants for a sales manager job might be asked to provide the names and dates of past sales team(s) and the specific accomplishments of the team.

WABs and BIBs have been used in a variety of settings for many types of jobs. WABs are used primarily for clerical and sales jobs. BIBs have been used successfully in the military and the insurance industry with an average validity of .35. Many insurance companies, for example, use a very lengthy BIB to screen their applicants. Check out www.e-Selex.com for an online biodata testing service.

The **accomplishment record** is an approach similar to a BIB. Job candidates are asked to write examples of their actual accomplishments, illustrating how they had mastered job-related problems or challenges. Obviously, the problems or challenges should be compatible with the problems or challenges facing the organization. The applicant writes these accomplishments for each of the major components of the job. For example, in a search for a new business school dean, applicants were asked to cite a fund-raising project they had successfully organized. HRM specialists evaluate these accomplishments for their predictive value or importance for the job to be filled. Accomplishment records are particularly effective for managerial, professional, and executive jobs.[7] In general, research indicates that methods such as BIBs and accomplishment records are more valid as predictors of future success than credentials or crude measures of job experience. For example, having an MBA versus only a bachelor's degree is not a particularly valid predictor of successful management performance. What an applicant has accomplished in past jobs or assignments is a more valid approach to assessing managerial potential.

How Do You Derive WAB or BIB or Accomplishment Record Weights?

To derive the weights for WABs or BIBs, you ideally need a large (at least 100) representative sample of application or biographical data and criterion data (e.g., job tenure and/or performance) of the employees who have occupied the position under study. You then can correlate responses to individual parts of the instrument with the criterion data. If effective and ineffective (or long-tenure versus short-tenure) employees responded to an item differently, responses to this item would then be given different weights, depending on the magnitude of the relationship. Weights for the accomplishment record are usually derived by expert judgment for various problems or challenges.

Research supports the use of WABs, BIBs, and the accomplishment record in selection. The development of the scoring system requires sufficient data and some research expertise, but it is worthwhile because the resulting decisions are often superior to those typically made based on a subjective interpretation of application blank information. What if you can't do the empirical validation study? Might you still get better results using a uniform weighted system, in which the weights are based on expert judgment? Yes. This approach is superior to one in which there is no uniform weighting system and each application blank or résumé is evaluated in a more holistic manner by whoever is evaluating it.

REFERENCE CHECKS AND BACKGROUND CHECKS

The vast majority of employers now conduct background checks on job applicants. The goal is to gain insight about the potential employee from people who have had previous experience with him or her. An important role of the background check is to simply verify the information provided by the applicant regarding previous employment and experience. This is a good practice, considering research indicates that between 20 and 25 percent of job applications include at least one fabrication.[8]

Many organizations are now "Googling" applicants' names and searching Facebook and MySpace for information about job candidates as part of a preliminary background check. Over a third of executive recruiters indicated in a recent survey that they eliminated job candidates based only on information that they found based on web searches of the candidates' "digital dossier." A great deal of this information is attributable social networking sites such as Facebook and LinkedIn.[9] In some states, administrators hiring teachers routinely search the web for potentially embarrassing (or worse) material. In some states, teachers have been removed for risqué web pages and videos. "I know for a fact that when a superintendent in Missouri was interviewing potential teachers last year, he would ask, 'Do you have a Facebook or MySpace page?' " said Todd Fuller, a spokesman for the Missouri State Teachers Association. The association is now warning its members to audit their web pages. "If the candidate said yes, then the superintendent would say, 'I've got my computer up right now. Let's take a look.' " The largely unregulated background check industry may be one of the fastest growing (and most profitable) of all HR areas today. These specialty firms often compile "digital dossiers" on individuals based on many sources, including web searches, interviews with past employers and co-workers, criminal and driving histories, and credit ratings.[10] Obviously, people need to closely monitor their web "presence" or "digital footprint" and exercise as much caution as possible to avoid future incriminating (or embarrassing) information.

Fear of **negligent hiring** lawsuits is a related reason that employers do reference and background checks. **A negligent hiring lawsuit is directed at an organization accused of hiring incompetent (or dangerous) employees.** Lawsuits for negligent hiring attempt to hold an organization responsible for the behavior of employees when there is little or no attempt by the organization to assess critical characteristics of those who are hired. There may be no limit to the liability an employer can face for this negligence. One health management organization was sued for $10 million when a patient under the care of a psychologist was committed to a psychiatric institution and it was later revealed that the psychologist was unlicensed and had lied about his previous experience.

Organizations also conduct reference checks to assess the potential success of the candidate for the new job. Reference checks provide information about a candidate's past performance and are also used to assess the accuracy of information provided by candidates. However, HR professionals should be warned: lawsuits have engendered a reluctance on the part of evaluators to provide anything other than a statement as to when a person was employed and in what capacity. These lawsuits have been directed at previous employers for defamation of character, fraud, and intentional infliction of emotional distress. One jury awarded a man $238,000 for defamation of character because a past employer erroneously reported that "he was late most of the time, regularly missed two days a week."[11] This legal hurdle has prompted many organizations to stop employees from providing any information about former employees other than dates of employment and jobs. Turnaround is fair play—at least litigiously. **Organizations are being sued and held liable if they do not give accurate information about a former employee when another company makes such a request.** At least one web-based company will check on what references say about you. At Badreferences.Com, for $87.95, you can receive a reference report from former employers, contractors, even professors. For more money, the same company will help prepare a "cease and desist" order and, for $120 per hour, provide court testimony on your behalf.

The bottom line appears simple: Tell the truth about former employees. There are laws in several states that provide protection for employers and former managers who provide candid and valid evaluations of former employees.

What Is the Validity of Reference Checks?

One of the problems with letters of reference is that they are almost always very positive. While there is some validity, it is low in general (.20–.30 range). One approach to getting more useful (and valid) distinctions among applicants is to construct a "letter of reference" or recommendation that is essentially a performance appraisal form. One can construct a rating form and request that the evaluator indicate the extent to which the candidate was effective in performing a list of job tasks. This approach offers the added advantage of deriving comparable data for both internal and external job candidates, since the performance appraisal, or reference data, can be completed for both internal and external candidates. One study found that reference checks significantly predicted subsequent supervisory ratings (0.36) when they were conducted in a structured and telephone-based format.[12] With this approach, both internal and external evaluators must evaluate performances on the tasks that are most important for the position to be filled.

An alternative method asks the evaluator to rate the extent of job-related knowledge, skill, ability, or competencies of a candidate. These ratings can then be weighted by experts based on the relative importance of the KASOCs or competencies for the position to be filled. This approach makes good sense whenever past performance is a strong predictor of future performance. For example, when selecting a manager from a pool of current or former managers, a candidate's past performance as a manager is important. Performance appraisals or promotability ratings, particularly those provided by peers, are a valid source of information about job candidates. However, promotability ratings made by managers are not as valid as other potential sources of information about candidates, such as scores on GMA or performance tests, and assessment centers. The validity of reference checking can be enhanced by gathering information from a larger number of references (10 to 12 if possible) and obtaining this information from sources other than those recommended by the job candidates.[13]

Employers should do their utmost to obtain accurate reference information about external candidates despite the difficulties. If for no other reason, a good-faith effort to obtain verification of employment history can make it possible for a company to avoid (or win) negligent hiring lawsuits.

What Are the Legal Implications of Doing Background Checks and Reference Checks on Job Candidates?

Employers often request consumer reports or more detailed "investigative consumer reports" (ICVs) from a consumer credit service as a part of the background check. If they do this, employers need to be aware of state laws related to background checks and the **Fair Credit Reporting Act (FCRA),** a federal law that regulates how such agencies provide information about consumers. State laws vary considerably on background checks. Experts maintain that it is legally safest to comply with the laws of the states where the job candidate resides, where the reporting agency is incorporated, and where the employer has its principal place of business. In general, in order to abide by the FCRA or state law, four steps must be followed by the employer: (1) Give the job candidate investigated a notice in writing that you may request an investigative report, and obtain a signed consent form; (2) provide a summary of rights under federal law (individuals must request a copy); (3) certify to the investigative company that you will comply with federal and state laws by signing a form it should provide; and (4) provide a copy of the report in a letter to the person investigated if a copy has been requested or if an adverse action is taken based on information in the report.

White-collar crime, including employee theft and fraud, is an increasingly serious and costly problem for organizations. One bad hire could wipe out a small business. Enter Ken Springer, a former FBI agent, and now the president of Corporate Resolutions, a fast-growing personnel investigation company with offices in New York, London, Boston, Miami, and Hong Kong. Many of Springer's clients are private equity firms that request management background checks at companies the equity firms are evaluating for possible purchase. Springer also does prescreening for management and executive positions.

Springer's major recommendation is to carefully screen all potential employees (because even entry-level employees can do major damage to an organization) and to carefully research and verify all information on the résumés. He believes that if a single lie is detected, the applicant should be rejected. In addition, Springer says to be wary of claims that

are difficult to verify, to carefully research all gaps in applicants' employment histories and vague descriptions of what they did, and to require and contact at least three references to verify as much information as possible. Springer also recommends that after verifying all facts in a job candidate's résumé, a thorough background check should be done.

Among other companies doing basic job candidate screening, with prices ranging from $100 to $400, are Taleo, Automatic Data Processing, HireRight, and National Applicant Screening. Google "employment screening" and you'll find numerous other companies doing preemployment screening and background checks for employers. It is advisable for employers to consult with the National Association of Professional Background Screeners (NAPBS) regarding firms to use for background and reference checks. The NAPBS was founded to promote ethical business practices, to comply with the **Fair Credit Reporting Act,** and to foster awareness of issues related to consumer protection and privacy rights within the background screening industry.

PERSONNEL TESTING

GMA tests are valid for virtually all jobs

Many organizations use general mental ability (GMA) (also known as cognitive ability tests) to screen applicants, bolstered by considerable research indicating that GMA tests are valid for virtually all jobs in the U.S. economy. The dilemma facing organizations is this: While GMA tests have been shown to be valid predictors of job performance, they can create legal problems because minorities tend to score lower. GMA tests are ideal for jobs if considerable learning or training on the job is required and where a more "job-related" knowledge-based test is inappropriate or unavailable.[14]

Corporate America also is increasing its use of various forms of personality or motivational testing—in part due to the body of evidence supporting the use of certain methods, concern over employee theft, the outlawing of the polygraph test, and potential corporate liability for the behavior of its employees. Domino's Pizza settled a lawsuit in which one of its delivery personnel was involved in a fatal accident. The driver had a long and disturbing psychiatric history and terrible driving record before he was hired.

The paper-and-pencil and online tests most frequently used today for employment purposes are GMA tests. These tests attempt to measure the verbal, quantitative, mechanical, or sensory capabilities in job applicants. You are probably familiar with these "high stakes" cognitive ability tests: the Scholastic Aptitude Test (SAT), the American College Test (ACT), the Graduate Management Admissions Test (GMAT), the Graduate Record Examination (GRE), and the Law School Admissions Test (LSAT).

Cognitive ability tests, most of which are administered in a paper-and-pencil or computerized format under standardized conditions of test administration, are controversial. On average, African Americans and Hispanics score lower than Whites on virtually all of these tests; thus, use of these tests for selection purposes can cause legal problems and difficulties for an organization seeking greater diversity in its workforce. The critical issue of test score differences as a function of ethnicity is discussed later in the chapter. Let's begin with a definition of GMA testing and provide brief descriptions of some of the most popular tests. Next, the validity evidence for these tests is reviewed.

What Is a Cognitive (or General Mental) Ability Test?

Cognitive ability or general mental ability (GMA) tests measure one's aptitude or mental capacity to acquire knowledge based on the accumulation of learning from all possible sources. Standardized tests of GMA are based on research that has focused on understanding individuals' ability to reason, plan, solve problems, think abstractly, learn and adapt, and process and comprehend complex ideas and information.

Such tests should be distinguished from **achievement tests,** which attempt to measure the effects of knowledge obtained in a standardized environment (e.g., your final exam in this course could be considered a form of achievement test). Cognitive ability or GMA tests are typically used to predict future performance. The SAT and ACT, for example, were developed to measure ability to master college-level material. Having made this

distinction between achievement tests and cognitive ability tests, however, in practice there isn't a clear distinction between these two classes of tests. Achievement tests can be used to predict future behavior, and all tests measure some degree of accumulated knowledge. **Knowledge-based tests** assess a sample of what is required on the job. If you are hiring a computer programmer, a cognitive ability test score might predict who will learn to be a computer programmer; but a better approach is an assessment of actual programming knowledge. Knowledge-based tests are easier to defend in terms of job relatedness and are quite valid (.48) and recommended for identifying those job candidates who can be highly effective the very first day of work (i.e., no training on the critical knowledge of the job required). However, knowledge tests can be expensive to develop.[15]

There are hundreds of GMA tests available. In addition to the "high stakes" tests, some of the most frequently used tests are the **Wechsler Adult Intelligence Scale, the Wonderlic Personnel Test, and the Armed Services Vocational Aptitude Battery.** In addition, many of the largest U.S. companies have developed their own battery of cognitive ability tests. AT&T evaluates applicants for any of its nonsupervisory positions on the basis of scores on one or more of its 16 mental ability subtests. McClachy, the communications giant, has a battery of 10 mental ability tests that are weighted differently for different jobs.

The **Wechsler Adult Intelligence Scale** is one of the most valid and heavily researched of all tests. A valid and more practical test is the **Wonderlic Personnel Test.** The publisher of this test, first copyrighted in 1938, has data from more than 3 million applicants. The Wonderlic consists of 50 questions covering a variety of areas, including mathematics, vocabulary, spatial relations, perceptual speed, analogies, and miscellaneous topics. Here is an example of a typical mathematics question: "A watch lost 1 minute 18 seconds in 39 days. How many seconds did it lose per day?" A typical vocabulary question might be phrased as follows: "Usual is the opposite of: a. rare, b. habitual, c. regular, d. stanch, e. always." An item that assesses ability in spatial relations would require the test taker to choose among five figures to form depicted shapes. Applicants have 12 minutes to complete the 50 items. The Wonderlic will cost an employer from $1.50 to $3.50 per applicant depending on whether the employer scores the test. The Wonderlic is used by the National Football League to provide data for potential draft picks (the average score of draftees is one point below the national population).[16]

The Wonderlic and the NFL

You may remember the Wonderlic from the discussion of the Supreme Court rulings in *Griggs v. Duke Power* (discussed in Chapter 3) and *Albemarle v. Moody.* In *Griggs,* scores on the Wonderlic had an adverse impact against African Americans (a greater proportion of African Americans failed the test than did whites), and *Duke Power* did not show that the test was job related. Despite early courtroom setbacks and a decrease in use following the *Griggs* decision, according to the test's publisher, the use of the Wonderlic has increased in recent years.

Current interest in cognitive ability tests was spurred by the research on **validity generalization,** which strongly supported the validity of these tests for virtually all jobs and projected substantial increases in utility for organizations that use the tests. Scores on GMA tests are strongly related to success in occupational training in both civilian and military jobs, with meta-analytic estimates ranging from the high .30s to .70s and averaging around .50. GMA scores are also related to overall job performance, objective leadership effectiveness, and assessments of creativity. The strength of the relationship between test scores and performance increases as training and jobs become more cognitively complex and mentally challenging. Validities also tend to be even higher for jobs that are dynamic, are fast changing, and require adaptability. Differences in GMA and in specific GMA ability patterns also predict differences in educational, occupational, and creative outcomes years later; that is, the relationships among an individual's math, verbal, and spatial abilities also predict lead outcomes in education, job performance, and creative endeavors 10 or more years later. Also, a convincing argument can be made that the validities for most employment selection methods are higher than previously thought. Using an appropriate statistical adjustment, increases in validity estimates were found to be greater for GMA than for self-report personality measures. In addition, the incremental validity of the personality measures over that provided by GMA scores alone was found to be smaller (but still significant) than previously estimated in past studies.[17]

GMA tests more valid for more complex jobs

Despite abundant research indicating the importance of GMA for complex jobs, it is interesting to note that over half of the top executive MBA programs, as rated by *BusinessWeek* magazine in 2005, had actually dropped the GMAT (General Management Admissions Test) for admissions to their programs. Also, according to one study after controlling for GMA, the MBA degree itself may not be a good predictor of long-term executive success.[18]

Figure 6-4 presents some myths regarding the use and interpretation of GMA tests. One of the more popular myths about GMA is that once a person reaches a certain threshold of GMA (e.g., a score on a GMA test), then differences on GMA do not matter; that is, these differences are not related to better performance. For example, Malcolm Gladwell writes in his best seller *Outliers: The Story of Success* that "The relationship between success and IQ works only up to a point. Once someone has an IQ of somewhere around 120, having additional IQ points doesn't seem to translate into any measurable real-world advantage."[19] In fact, abundant research indicates that even at the top 1 percent of GMA, a higher level of GMA is related to higher performance.[20]

What Are Tests of Specific Mental Abilities?

A variety of tests have also been developed to measure specific abilities, including specific cognitive abilities or aptitudes such as verbal comprehension, numerical reasoning, and verbal fluency, as well as tests assessing mechanical and clerical ability and physical or psychomotor ability, including coordination and sensory skills. The most widely used mechanical ability test is the **Bennett Mechanical Comprehension Test (BMCT).** First developed in the 1940s, the BMCT consists mainly of pictures depicting mechanical situations with questions pertaining to the situations. The respondent describes relationships between physical forces and mechanical issues. The BMCT is particularly effective in the prediction of success in mechanically oriented jobs.

While there are several tests available for the assessment of clerical ability, the most popular is the **Minnesota Clerical Test (MCT).** The MCT requires test takers to quickly compare either names or numbers and to indicate pairs that are the same. The name

Figure 6-4 Myths about the Usefulness of General Mental Ability

1. There is no relationship with important outcomes such as creativity or leadership.

   FINDING: Scores on GMA tests are strongly related to success in academic domains, job for both civilian and military jobs with meta-analytic estimates from the high .30s to .70s. GMA scores also predict important outcomes in all jobs including overall job performance, leadership effectiveness, and assessments of creativity.

2. There is predictive bias when using GMA tests.

   FINDING: Research on the fairness of ability tests has drawn the conclusion that tests are not biased against women and minority groups. More informal hiring practices are much more likely to be biased.

3. There is a lack of predictive independence from a test takers' socioeconomic status (SES).

   FINDING: SES is related to test scores but to only a modest degree. SES variables do not eliminate the predictive power of GMA tests. SES does not explain the relationship between test scores and subsequent performance.

4. There are thresholds beyond which scores cease to matter.

   FINDING: More ability is associated with greater performance (e.g., College GPA is linearly related to SAT test scores across the entire range of scores). Correlations between supervisors' ratings of employees' job performance are linearly related to GMA.

5. Other characteristics, especially personality, are more valid than GMA.

   FINDING: Measures of personality, habits, and attitudes can produce useful incremental validity in predicting performance but validities of GMA (versus self-report measures of non-cognitive factors) are higher.

Adapted from the following sources: Robertson, K. F., Smeets, S., Lubinski, D., & Benbow, C. P. (2010). Beyond the threshold hypothesis: Even among the gifted and top math/science graduate students, cognitive abilities, vocational interests, and lifestyle preferences matter for career choice, performance, and persistence. *Current Directions in Psychological Science, 19,* 346–351; Connelly, B. S., & Ones, D. S. (2010). Another perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136,* 1092–1122; Coward, W. M., & Sackett, P. R. (1990). Linearity of ability performance relationships: A reconfirmation. *Journal of Applied Psychology, 75,* 297–300; Kuncel, N. R., & Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science, 19,* 339–345; Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate student success (supplementary material). *Science, 315,* 1080–1081; Ones, D. S., Viswesvaran, C., & Dilchert, S. (2005). Cognitive ability in personnel selection decisions. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), *The Blackwell handbook of personnel selection.* Oxford, UK: Blackwell; Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High stakes testing in higher education and employment. *American Psychologist, 63,* 215–227; Sackett, P. R., Kuncel, N. R., Arneson, J., Cooper, S. R., & Waters, S. (2009). Socio-economic status and the relationship between admissions tests and post-secondary academic performance. *Psychological Bulletin, 135,* 1–22.

comparison part of the test has been shown to be related to reading speed and spelling accuracy, while the number comparison is related to arithmetic ability.

Research on the use of specific abilities versus GMA favors the use of the GMA in the prediction of training success and (probably) job performance as well. A meta-analysis concluded that "weighted combinations of specific aptitudes tests, including those that give greater weight to certain tests because they seem more relevant to the training at hand, are unnecessary at best. At worst, the use of such tailored tests may lead to a reduction in validity."[21]

Are There Racial Differences in Test Performance?

Griggs v. Duke Power

Many organizations discontinued the use of cognitive ability tests because of the Supreme Court ruling in *Griggs.* Despite fairly strong evidence that the tests are valid and their increased use by U.S. businesses, the details of the *Griggs* case illustrate the continuing problem with the use of such tests. The Duke Power Company required new employees either to have a high school diploma or to pass the Wonderlic Personnel Test and the Bennett Mechanical Comprehension Test. Fifty-eight percent of whites who took the tests passed, while only 6 percent of African Americans passed. According to the Supreme Court, the Duke Power Company was unable to provide sufficient evidence to support the job relatedness of the tests or the business necessity for their use. Accordingly, based on the "disparate impact" theory of discrimination, the Supreme Court ruled that the company had discriminated against African Americans under Title VII of the 1964 Civil Rights Act. As discussed in Chapter 3, the rationale for the Supreme Court's decision gave rise to the theory of disparate impact.

The statistical data presented in the *Griggs* case are not unusual. African Americans, on average, score significantly lower than whites on GMA tests; Hispanics, on average, fall about midway between average African American and white scores.[22] Thus, under the disparate impact theory of discrimination, plaintiffs are likely to establish adverse impact based on the proportion of African Americans versus whites who pass such tests. If the *Griggs* case wasn't enough, the 1975 Supreme Court ruling in *Albemarle Paper Company v. Moody* probably convinced many organizations that the use of cognitive ability tests was too risky. In *Albemarle,* the Court applied detailed guidelines to which the defendant had to conform in order to establish the job relatedness of any selection procedure (or job specification) that caused adverse impact in staffing decisions. The *Uniform Guidelines in Employee Selection Procedures,* as issued by the Equal Employment Opportunity Commission, also established rigorous and potentially costly methods to be followed by an organization to support the job relatedness of a test if adverse impact should result.

Some major questions remain regarding the validity generalization results for cognitive ability tests: Are these tests the most valid method of personnel selection across all job situations or are other methods, such as biographical data and personality tests, more valid for some jobs that were not the focus of previous research? Are there procedures that can make more accurate predictions than cognitive ability tests for some job situations? Are cognitive ability tests the best predictors of sales success, for example? (Remember the Unabomber? He had a near perfect SAT score and a PhD in math from the University of Michigan. How would he do in sales?) Another issue is the extent to which validity can be inferred for jobs involving bilingual skills. Would the Wonderlic administered in English have strong validity for a job, such as a customs agent, requiring the worker to speak in two or more languages? Bilingual job specifications are increasing in the United States. Invoking the "validity generalization" argument for this type of job based on research involving only the use of English is somewhat dubious. The validity of such tests to predict performance for these jobs is probably not as strong as .5.

Another issue concerns the extent to which other measures can enhance predictions beyond what cognitive ability tests can predict. Generally, human performance is thought to be a function of a person's ability, motivation, and personality. The average validity of cognitive ability tests is about 0.50. This means that 25 percent of the variability in the criterion measure (e.g., performance) can be accounted for by the predictor, or the test. That leaves 75 percent unaccounted for. Industrial psychologists think the answer lies in measures of one's motivation to perform, personality, or the compatibility of a person's job preferences with actual job characteristics.

Would a combination of methods—perhaps a cognitive ability test and a personality or motivational test—result in significantly better prediction than the GMA test alone? Research indicates that a combination of cognitive and non-cognitive assessments (e.g, measures of a job candidate's motivation or personality) may lead to a more comprehensive assessment of an individual and potentially higher validity than any method by itself.[23] Motivational or personality assessments through tests, questionnaires, interviews or other methods add what is known as **incremental validity** in the prediction of job performance. In general, GMA and job knowledge tests are highly valid but additional (and valid) tools can improve validity of personnel decisions and also have the potential to reduce adverse impact. In general, measures of personality, work habits or preferences, and attitudes demonstrate low to zero correlations with GMA and, therefore, produce very useful incremental validity in predicting performance across most jobs.[24] Accordingly the use of other selection methods that address the non-cognitive components of human performance, in addition to a GMA/cognitive ability or knowledge-based test, can help an organization make better decisions (and with less adverse impact). These measures are discussed shortly.

What is incremental validity?

Why Do Minorities Score Lower than Whites on GMA Tests?

This question has interested researchers for years, yet there appears to be no clear answer. Most experts now generally take the view that these differences are not created by the tests but are most related to inferior educational experiences. But the problem is not a defect or deficiency in the tests per se. The critical issue for HRM experts is not how to modify the test itself, but how to use the test in the most effective way. A panel of the **National Academy of Sciences** concluded that cognitive ability tests have limited but real ability to predict how well job applicants will perform, and these tests predict minority group performance as well as they predict the future performance of nonminorities. In other words, the tests themselves are not to blame for differences in scores. Obviously, the dilemma for organizations is the potential conflict in promoting diversity while at the same time using valid selection methods that have the potential for causing adverse impact. As one recent review concluded, "Although the evidence indicates that the group differences reflected by standardized cognitive tests are not caused by the tests themselves, we need to decide how to address the causes of group differences and wrestle with their consequences. We should continue to strive to further understand the nature and development of cognitive abilities and seek additional assessments that supplement cognitive ability test scores to improve decision-making accuracy."[25]

How Do Organizations Deal with Race Differences on Cognitive Ability Tests?

The use of top-down selection decisions based strictly on scores on cognitive ability tests is likely to result in adverse impact against minorities. One solution to this problem is to set a cutoff score on the test so as not to violate the 80 percent rule, which defines adverse impact. Scores above the cutoff score are then ignored and selection decisions are made on some other basis. The major disadvantage of this approach is that there will be a significant decline in the utility of a valid test because people could be hired who are at the lower end of the scoring continuum, making them less qualified than people at the upper end of the continuum who may not be selected. **Virtually all of the research on cognitive ability test validity indicates that the relationship between test scores and job performance is linear; that is, higher test scores go with higher performance and lower scores go with lower performance.** Thus, setting a low cutoff score and ignoring score differences above this point can result in the hiring of people who are less qualified. So, while use of a low cutoff score may enable an organization to comply with the 80 percent adverse impact rule, the test will lose considerable utility.

GMA has a linear relationship with performance

What is banding?

Another approach to dealing with potential adverse impact is to use a **banding** procedure that groups test scores based on data indicating that the bands of scores are not significantly different from one another. The decision maker then may select anyone from within this band of scores. Banding is not unlike grade distributions where scores from 92–100 percent all receive an "A," 82–91 receive a "B," and so on. Where banding can get contentious is when an organization invokes an argument that scores within a band are "equal" and then selection is made based on a protected class characteristic to promote diversity or as part

of an affirmative action program. Unfortunately, research shows that banding procedures have a big effect on adverse impact only when minority preference within a band is used for selection. This approach is controversial and may be illegal.[26]

The use of cognitive ability tests obviously presents a dilemma for organizations. Evidence indicates that such tests are valid predictors of job performance and academic performance and that validity is higher for jobs that are more complex (see again Figure 6-2). Employers that use such tests enjoy economic utility with greater productivity and considerable cost savings. However, selection decisions that are based solely on the scores of such tests will result in adverse impact against African Americans and Hispanics. Such adverse impact could entangle the organization in costly litigation and result in considerable public relations problems. If the organization chooses to avoid adverse impact, the question becomes one of either throwing out a test that has been shown to be useful in predicting job performance or keeping the test and somehow reducing or eliminating the level of adverse impact. But does such a policy leave a company open to reverse discrimination lawsuits by whites who were not selected for employment since their raw scores on the test were higher than scores obtained by some minorities who were hired? Many organizations, particularly in the public sector, have abandoned the use of cognitive ability tests in favor of other methods, such as interviews or performance tests, which result in less adverse impact and are more defensible in court. However, many other cities and municipalities have opted to keep such tests and then have employed some form of banding in the selection of their police and firefighters primarily in order to make personnel decisions that do not result in statistical adverse impact.

Researchers and practitioners are very interested in how to select the most effective candidates while meeting diversity goals and minimizing (or eliminating) adverse impact. There have been some criticisms of the tests themselves with suggestions to remove the "culturally biased" questions. However, research does not support this recommendation. Research also does not support dropping use of GMA or knowledge-based tests. While many approaches have been proposed and have been taken to reduce statistical adverse impact against minorities, research indicates that some recommendations can be made.

1. Target recruitment strategies toward "qualified" minorities.
2. Focus on predicting all aspects of job performance, including citizenship behavior, helping co-workers, teamwork, and counter-productive behavior.
3. Augment GMA test use with noncognitive methods such as personality tests, peer assessments, interviews, and job preference instruments.
4. Use tools with less adverse impact early in the process and GMA tests later providing the selection ratio is low.
5. Use accomplishment records, performance tests, or work samples in lieu of GMA tests.

What Are Physical or Psychomotor Tests?

Physical, psychomotor, and sensory/perceptual tests are classifications of ability tests used when the job requires particular abilities. Physical ability tests are designed to assess a candidate's physical attributes (e.g., muscular tension and power, muscular endurance, cardiovascular endurance, flexibility, balance, and coordination). Scores on physical ability tests have been linked to accidents and injuries, and the criterion-related validity for these tests is strong. One study found that railroad workers who failed a physical ability test were much more likely to suffer an injury at work. Psychomotor tests assess processes such as eye–hand coordination, arm–hand steadiness, and manual dexterity. Sensory/perceptual tests are designed to assess the extent to which an applicant can detect and recognize differences in environmental stimuli. These tests are ideal for jobs that require workers to edit or enter data at a high rate of speed and are also valid for the prediction of vigilant behavior. Recall our discussion earlier that Wackenhut Security was seeking more vigilant armed security guards. Researchers focused on tests that assessed this skill and found evidence that sensory/perceptual tests could predict this particular attribute.

As discussed in Chapter 3, based on the ADA and Title VII, the validity of physical ability tests has been under close scrutiny. For example, many Title VII lawsuits have been filed on behalf of female applicants applying for police and firefighter jobs who had failed some type of physical ability test that purports to assess physically demanding attributes of the job. In fact, the probability is high for adverse impact against women when a physical ability test is used to make selection decisions. For example, the strength tests will probably have adverse impact against women (almost two-thirds of all males score higher than the highest scoring female on muscular tension tests).[27] Job analysis data are clearly needed to establish this attribute as an essential element of a job and that such an attribute is stated in a job description.

Sensory ability testing concentrates on the measurement of hearing and sight acuity, reaction time, and psychomotor skills, such as eye and hand coordination. Such tests have been shown to be related to quantity and quality of work output and accident rates.[28]

What Is Personality/ Motivational/ Dispositional Testing?

While research supports the use of GMA tests for personnel selection, performance is a function of both ability and motivation. Scores on GMA or other ability or knowledge-based tests say little or nothing about a person's motivation or personality to do the job. We can all think of examples of very intelligent individuals who were unsuccessful in many situations (we're back to the Unabomber or perhaps you remember Bobby Fisher, the great but troubled chess player!). Most of us can remember a classmate who was very bright but received poor grades due to low motivation. The validity of GMA tests for predicting sales success is significant but low and we can definitely improve on prediction by using other assessment tools in addition to a GMA test.[29]

Most personnel selection programs attempt an informal or formal assessment of an applicant's personality, motivation, attitudes, or disposition through psychological testing, reference checks, or a job interview. Some of these so-called noncognitive assessments are based on scores from standardized tests, performance testing such as job simulations, or assessment centers. Others are more informal, derived from an interviewer's gut reaction or intuition. This section reviews the abundant literature on the measurement and prediction of motivation, dispositions, and personality characteristics using various forms of assessment. Without question, some approaches are more valid than others and some are not valid at all for use in staffing decisions.

There is an increased use of various types and formats for personality or motivational testing, including on-line assessment, video and telephone testing. There is also increasing evidence that many of these methods are valid predictors of job performance and other important criteria such as job tenure or turnover and **counterproductive work behavior (CWB)** such as employee theft, aberrant or disruptive behaviors, and interpersonal and organizational deviance.

Predicting counterproductive behavior

Some organizations place great weight on personality testing for employment decisions. A 2006 survey indicated that 35 percent of U.S. companies use personality tests for personnel selection.[30] The increase in usage may be partially a function of the trend toward more interdependent, team-based, and project-based organizations with an increased importance placed on the compatibility of the team members. Team members' personalities are clearly related to this compatibility. Research shows that certain traits can predict how people behave and perform in groups.[31] We'll review this literature after we define personality and describe some of the most popular tests that measure personality traits.

Although the criterion-related validity evidence made available to the public is rather limited, one of the most popular personality assessment tools is the **"Caliper Profile,"** developed by the Caliper Corporation (www.calipercorp.com). Its website claims 25,000 clients. BMW, Avis, and GMAC are among the companies that use the Caliper Profile to hire salespeople. The profile has also been used by numerous sports teams for player personnel issues such as potential trades and drafts. The Chicago Cubs, the Detroit Pistons, and the New York Islanders are among the sports teams that have used the profile for drafting and trade considerations (not exactly a ringing endorsement). Many companies have hired consultants to screen job candidates for their "**emotional intelligence**" (EI), probably influenced far less by sound research than by the popularity of the approach, the

plethora of consulting in this area, and the 1995 best-seller "Emotional Intelligence" by Daniel Goleman, who claimed that emotional intelligence is a stronger predictor of job performance than GMA (it isn't at least as reported in peer-reviewed, scholarly journals).[32] Try HayGroup.com for one of the most popular firms specializing in EI. Sears, IBM, and AT&T have used personality tests for years to select, place, and even promote employees. Many companies today use some form of personality test to screen applicants for risk factors related to possible counterproductive behavior.

There are literally thousands of personality tests and questionnaires available that purport to measure hundreds of different traits or characteristics. (Go to www.unl.edu/buros/ for a sample.) The basic categories of personality testing are reviewed next. Figure 6-5 presents a list of some of the most popular tests and methods.

Let's start with a definition of personality and provide brief descriptions of some of the more popular personality tests. The validity of the major personality tests is reviewed along with an overview of relevant legal and ethical issues. The section concludes with a description of some relatively new "noncognitive" tests that have shown potential as selection and placement devices.

What Is Personality?

While personality has been defined in many ways, the most widely accepted definition is that **personality** refers to an individual's consistent pattern of behavior. This consistent pattern is composed of psychological traits. While a plethora of traits have been labeled and defined, most academic researchers subscribe to a five-factor model (FFM) to describe personality.[33] These so-called Big Five personality factors are as follows: (1) **Emotional stability (**also known **as Neuroticism)**); (2) **Extraversion** (outgoing, sociable); (3) **Openness to experience** (imaginative, curious, experimenting); (4) **Agreeableness** (friendliness, cooperative vs. dominant); and (5) **Conscientiousness** (dependability, carefulness). There are several questionnaires or inventories that measure the FFM. (Try http://users.wmin.ac./UK/~buchant/ for a free online "Big Five" test.) There is research supporting the validity of the FFM in the prediction of a number of criteria (e.g., performance, sales, counterproductive behaviors) for a variety of jobs. This validity evidence is reviewed in a later section.

The "Big Five" or FFM

Two relatively new characterizations of personality are **Emotional Intelligence (EI)** and **Core Self-Evaluations (CSE).** A 2008 count found 57 consulting firms devoted primarily to EI and about 90 firms specializing in training or assessment of EI, 30 EI certification programs, and five EI "universities."[34] EI is considered to be a multidimensional form or subset of social intelligence or a form of social literacy. EI has been the object of criticism because of differences in definitions of the contruct and the claims of validity and incremental validity. One definition is that **EI is a set of abilities that enable individuals**

Figure 6-5
Some Examples of Personality/Dispositional/Motivational Tests

PROJECTIVE TECHNIQUES AND INSTRUMENTS
Thematic Apperception Test (TAT)
Miner Sentence Completion Scale (MSCS)
Graphology (handwriting analysis)
Rorschach Inkblot Test

SELF-REPORT INVENTORIES—EXAMPLES
The NEO-PI-R Personality Inventory (measures FFM and facets of each)
Personal Characteristics Inventory
DiSC Profile
Myers-Briggs Type Indicator
Minnesota Multiphasic Personality Inventory (MMPI)
California Personality Inventory (CPI)
Sixteen Personality Factors Questionnaire (16 PF)
Hogan Personality Inventory
Job Compatibility Questionnaire (JCQ)
Emotional Intelligence (e.g., EI Scale)
Core Self-Evaluations Scale (CSES)
Caliper Profile

to recognize and understand their own emotions and those of others in order to guide their thinking and behavior to help them cope with the environment. One review concluded that "we are still far from being at the point of rendering a decision as to the incremental value of EI for selection purposes."[35]

CSE is a broad and general personality trait composed of four heavily researched traits: (1) self-esteem (the overall value that one places on oneself as an individual); (2) self-efficacy (an evaluation of how well one can perform across situations); (3) neuroticism (the tendency to focus on the negative); and (4) locus of control (the extent to which one believes s/he has control over life's events). The core self-evaluation is a basic assessment of one's capability and potential.[36]

Incremental validity

There is some research that investigated the extent to which EI and CSE scores add incremental validity in the prediction of performance beyond the Big Five or other selection tools. In general, this research indicates useful incremental validity for both the EI construct and CSE..[37]

How Do We Measure Personality?

Personality tests can be sorted into two broad categories: projective tests and self-report inventories. Of course, we also can use the interview and data from other sources such as peer ratings or references as a means for assessing personality characteristics or competencies as well. **Projective tests** have many common characteristics, the most significant of which is that the purpose and scoring procedure of the tests are disguised from the test taker.[38]

Much concern has been expressed about the ability of job candidates to fake a self-report personality inventory in order to provide a more favorable impression to an employer. Projective tests make it very difficult to fake responses since the test taker has little or no idea what a favorable response is. One of the most famous projective tests is the **Rorschach Inkblot Test,** which presents a series of inkblots to respondents who must then tell a story of what they see in each one.

While numerous projective tests exist, the **Miner Sentence Completion Scale (MSCS)** is one of the few such tests specifically designed for use in the employment setting and with some validity evidence to back its use. Its aim is to measure managers' motivation to manage others.[39] The test appears to work. The test consists of 40 incomplete sentences, such as "My family doctor . . . ," "Playing golf . . . ," and "Dictating letters. . . ." The test taker is instructed to complete each sentence. According to the developer of these tests, the way in which an applicant completes the sentences reflects his or her motivation along seven areas. These areas are capacity to deal with authority figures, dealing with competitive games, handling competitive situations, assertiveness, motivation to direct others, motivation to stand out in a group, and desire to perform day-to-day administrative tasks. On the downside, the MSCS is expensive and there isn't a great deal of validity evidence to support its use.

Another projective test that has been used occasionally for employment purposes is the **Thematic Apperception Test,** or **TAT,** a test that typically consists of 31 pictures that depict a variety of social and interpersonal situations. The subject is asked to tell a story about each picture to the examiner. Of the 31 pictures, 10 are gender-specific while 21 others can be used with adults of either sex. Test takers are asked to describe who the people are in each picture and what is happening in the situation, which is clearly open to interpretation. The test taker then "projects" the outcome of the situation. Although a variety of scoring systems have been developed for interpreting a test taker's responses, one of the most popular approaches involves rating the responses with regard to the test taker's need for power (i.e., the need to control and influence others), achievement (i.e., the need to be successful), and affiliation (i.e., the need for emotional relationships). Like the MSCS, the TAT has been used for managerial selection and the limited research indicates some validity as a predictor of managerial and entrepreneurial success. AT&T has been using the TAT for years as a part of its assessment center to identify high-potential managerial talent.[40]

One form of projective test (discussed earlier) that has received considerable attention recently is **graphology,** or handwriting analysis. With this approach, a sample of your handwriting is mailed to a graphologist who (for anywhere from $10 to $50) provides an assessment of your intelligence, creativity, emotional stability, negotiation skills,

problem-solving skills, and numerous other personal attributes. According to some writers, graphology is used extensively in Europe as a hiring tool. *The Wall Street Journal* and *Inc.* magazine have reported an increase in the use of the method in the United States since 1989. One handwriting analysis company *reports that* "With the government pulling the plug on the polygraph, and employers clamming up on job references and liabilities from negligent hiring, it is one alternative managers are exploring in an effort to know whom they are hiring."[41] While the use of the method may be increasing, there is no compelling evidence that the method does anything but provide an assessment of penmanship. The only peer-reviewed and published studies on the validity of graphology have found no validity for the approach.[42]

Self-Report Personality Inventories

Self-report inventories, which purport to measure personality or motivation with the respondent knowing the purpose and/or the scoring procedure of the test, are much more common than projective techniques. Some instruments screen applicants for aberrant or deviant behavior (e.g., the MMPI), others attempt to identify potentially high performers, and others, particularly more recently developed tests, are directed at specific criteria such as employee theft, job tenure/turnover, accident proneness, or customer orientation.

Self-report inventories typically consist of a series of short statements concerning one's behavior, thoughts, emotions, attitudes, past experiences, preferences, or characteristics. The test taker responds to each statement using a standardized rating scale. During the testing, respondents may be asked to indicate the extent to which they are "happy" or "sad," "like to work in groups," "prefer working alone," and so forth.

One of the most popular and respected personality tests is the **Minnesota Multiphasic Personality Inventory (MMPI).** The MMPI is used extensively for jobs that concern the public safety or welfare, including positions in law enforcement, security, and nuclear power plants. The MMPI is designed to identify pathological problems in respondents, not to predict job effectiveness. The revised version of the MMPI consists of 566 statements (e.g., "I am fearful of going crazy"; "I am shy"; "Sometimes evil spirits control my actions"; "In walking, I am very careful to step over sidewalk cracks"; "Much of the time, my head seems to hurt all over"). Respondents indicate whether such statements are true, false, or they cannot say. The MMPI reveals scores on 10 clinical scales, including depression, hysteria, paranoia, and schizophrenia, as well as four "validity" scales, which enable the interpreter to assess the credibility or truthfulness of the answers. Millions of people from at least 46 different countries, from psychotics to Russian cosmonauts, have struggled through the strange questions.[43]

Litigation related to **negligent hiring** often focuses on whether an organization properly screened job applicants. For example, failure to use the MMPI (or ignoring MMPI results) in filling public-safety jobs has been cited in legal arguments as an indication of negligent hiring—although not always persuasively. Unfortunately, some companies are damned if they do and damned if they don't. Target stores negotiated an out-of-court settlement based on a claim of invasion of privacy made by a California job candidate who objected to a few questions on the MMPI being used to hire armed guards. Had one of the armed guards who was hired used his or her weapon inappropriately (and Target had not used the MMPI), Target could have been slapped with a negligent hiring lawsuit.

Another popular instrument is the **16 Personality Factors Questionnaire (16PF),** which provides scores on the factors of the FFM, plus others. In addition to predicting performance, the test is used to screen applicants for counterproductive work behavior, such as potential substance abuse or employee theft. AMC Theaters, C&S Corporation of Georgia, and the U.S. State Department are among the many organizations that use the 16PF to screen job candidates. An advantage of the 16PF over other self-report inventories is that one of the 16PF factors reveals a reliable and valid measure of GMA as well as scores on the Big Five factors and "Big-Five subfactors" or facets (discussed later).[44]

NEO-PI-R (FFM)

Although there are many instruments available, the **NEO Personality Inventory** is one of the most reliable and valid measures of the FFM.[45] Another popular instrument for employee development and team diagnostics rather than for selection purposes is the

Myers-Briggs

Myers-Briggs Type Indicator (MBTI).[46]

What Is the Validity of Personality Tests?

Potentially useful personality tests exist among a great number of bad ones, making it difficult to derive general comments regarding their validity. Some instruments and the factors they measure have shown adequate (and useful) validity while others show little or no validity for employment decisions. In general, the validity is lower for self-report personality inventories than for cognitive ability tests. However, personality assessments from others (e.g., peers) appears to have strong validity.[47]

The one projective instrument with a fairly good but limited track record for selecting managers is the MSCS. A review of 26 studies involving the MSCS found an average validity coefficient of .35.[48] However, almost all of this research was conducted by the test publisher and not published in peer-reviewed journals.

MSCS validity = .35

Conscientiousness and emotional stability have validity for all jobs

The latest review of the FFM found that self-reported **Conscientiousness** and **Emotional Stability** had useful predictive validity across all jobs but that Conscientiousness had the highest validity (.31). **Extraversion, Agreeableness, and Openness to Experience** had useful predictive validity but for only certain types of jobs.[49] For example, extraverts are more effective in jobs with a strong social component, such as sales and management. Extraversion is not a predictor of job success for jobs that do not have a strong social component (e.g., technical or quantitative work). More Agreeable workers are more effective team members. People with high scores on Openness to Experience are more receptive to new training and do well in fast-changing jobs that require innovative or creative thinking. Research also supports the use of the FFM in an effort to reduce absenteeism among workers.

Extraversion has validity for managerial jobs

Another meta-analysis that focused on the relationship between the FFM and leadership effectiveness concluded with the following (corrected) correlations: Extraversion (.31), Emotional Stability (.24), Agreeableness (.10), Conscientiousness (.28), and Openness to Experience (.10). Experts in managerial selection concluded that the "combination of these meta-analytic results firmly supports the use of personality scales in managerial selection."[50]

There is also evidence that criterion-related validities change significantly over time. A study of an entire European country's 1997 cohort of medical students found that over time, Extraversion, Openness, and Conscientiousness scores showed increases in operational validity in the prediction of medical school grade point averages. The authors report that while there may not be any advantages to being open and extraverted for early academic performance, these traits gain importance for later academic performance, probably when applied practice plays a greater role in the curriculum. Conscientiousness was found to be an increasing asset for medical students with validities going from .18 to .45. They concluded that in assessing the utility of personality measures, relying on early criteria (e.g., first year GPA) might underestimate the predictive value of personality variables.[51]

A particular combination of FFM factors can also predict important criteria more successfully than the factors in isolation. For example, the combination of Emotional Stability (neuroticism) and Extraversion, describing a "happy" person, is a better predictor of job performance in health care than either trait in isolation.[52] Another study found that the combination of highly Agreeable and low to moderately Conscientious managers were the least effective managers for evaluating and developing employees.[53] **Research involving the FFM and managerial performance shows that Conscientiousness (.28), Extraversion (.21), and Emotional Stability (.19) are useful predictors of managerial success and that scores on these three factors should be used to select managers**.[54]

Recent research also suggests that we might do a better job predicting performance with more narrowly defined traits or subfactors that define a broader trait such as one from the FFM. A meta-analysis found that narrow traits underlying the Conscientiousness (C) factor from the FFM provided incremental predictive validity above and beyond the global Conscientiousness measure. Thus, the subfactors of C (achievement, dependability, order, cautiousness) helped improve the prediction of job performance. There is also evidence that underlying narrow traits of Extraversion might help enhance prediction for certain criterion measures for sales jobs. However, the degree to which the subfactors contribute to prediction depends on the particular performance criterion and the particular occupation under study. For example, in the meta-analysis, a subfactor of Extraversion, called

Use FFM subfactors to increase validity

potency, was a more valid predictor of overall job proficiency, sales effectiveness, and irresponsible work behavior, while another subfactor, **affiliation,** was a stronger predictor of technical proficiency.[55]

Why is the validity of personality inventories low (relative to measures of GMA)? Although not supported by research, most people think that an employee's motivation or personality or emotional "intelligence" is much more important for job performance than is the employee's GMA. So why is the validity of GMA so much stronger than the validities for the noncognitive types of inventories? Experts have given a number of explanations for the low (but useful) validity of personality and motivational tests in the employment context. First, and most obvious, applicants can "fake" personality tests so their personality as reflected on the tests is compatible with the requirements of the job. In essence, in an earnest effort to gain employment, many applicants will try to make responses on a self-report personality inventory that they at least think will make them look as favorable as possible to the prospective employer. (One cannot fake the SATs or the GMATs.) There is no question that applicant faking on most noncognitive measures occurs, but what is not clear is the extent to which faking reduces the validity of personality tests. Most researchers believe that the decrease in the predictive validity of personality measures due to faking is modest. Faking is apparently more problematic for self-report personality inventories (e.g., NEO Inventory) than for some alternative methods of assessing personality (i.e., peer assessments, structured interviews and assessment centers).[56] There appears to be less faking when trained interviewers are used to do assessments compared to self-report questionnaires.[57]

Effects of response faking

Second, experts have been critical of the research designs in validation work and contend that more carefully designed research (with larger sample sizes) would demonstrate higher validity for personality tests. While validities still lag behind that of GMA and other cognitive measures, the improved designs have shown practically useful (but still relatively low) validities for many noncognitive measures and particularly as "add-ons" to GMA or knowledge-based tests for incremental validity. Research shows that the weight given to particular personality factors (or combinations of factors) should derive from a careful job analysis or from criterion-related validation research.

Another explanation for the relatively low correlations reported for personality measures is that the assumptions were that, like for GMA, personality traits and performance have linear relationships with performance. In fact, a growing body of literature is accumulating that shows that **the relationship between certain personal attributes and effectiveness is curvilinear** (not linear). For example, individuals seen either as low in assertiveness or as high in assertiveness are generally appraised as less effective leaders than others who have high (but not too high) levels of assertiveness. The ideal score on this dimension is to be above average on assertiveness but only a little above average. Managers who score either very high on self-report measures of assertiveness or who are perceived as such by subordinates are not rated as effective by their subordinates compared to managers who are moderately high on this factor.[58] Recent research has also found that the FFM dimensions of Conscientiousness, Emotional Stability, and Agreeableness may also have curvilinear relationships with performance.[59]

Another possible explanation is that behavior is to a great extent determined situationally, making stable personality traits unpredictable for criteria such as job performance or employee turnover. Recall some of the examples of items from personality tests listed earlier in this chapter. Note that most of the examples are not specific to the workplace; in fact, most of them are quite general. Research in other areas has found that behavior is dependent on the situation. A person who is friendly outside of work might be less sociable in the work setting. In order to enhance predictability, some research indicates that personality assessment should involve "contexualizing" the frame of reference for completing a personality instrument for selection purposes. The use of a job-related frame of reference (e.g., "I pay close attention to details at work") has been found to show potential for the criterion-related validity of personality scales.[60]

Frame of reference personality assessment

Most experts recommend the use of more than one method (e.g., inventories, peer assessments, interviews) and more effort to link particular traits (or subfactors) with particular work criteria. Personality assessment could be more specific to the workplace

and target particular criterion measures of interest, such as job retention/turnover, counterproductive work behavior such as employee theft, attendance, or particular and important functions of a job (e.g., driving behavior, customer service). One study proposes that job performance can be broken down into three general domains: task performance (the essence of the job), citizenship performance (a good organizational co-worker), and counterproductive work behavior (theft, deviance). Cognitively loaded predictors such as GMA and knowledge-based tests are the strongest predictors of task performance while noncognitive predictors are the best predictors in the citizenship and counterproductive domains.[61]

One recent meta-analysis clearly established the superior validity of significant other ratings (e.g., peers) of personality characteristics compared to self-report measures.[62] For example, the predictive power of other ratings of Conscientiousness was found to be greater than that of self-ratings (.29 vs. .20). Ratings from peers (and others) had substantially higher predictive validities than self-report data and incremental to self-report data. Also, other ratings of Emotional Stability, Openness, and Agreeableness showed fairly strong validity for predicting job performance while the validity of self-ratings of these traits were negligible. The authors concluded that "These results suggest that other reports may indeed provide stronger validities for predicting job performance than do self-report measures." Note also that the true score validities expected from combining large numbers of other raters for rating Conscientiousness, Emotional Stability, Agreeableness, or Openness are extremely high (.55, .37, .31, and .45, respectively). Indeed, these considerably exceed validities for predicting job performance from personality ratings reported in any past, large-scale research. They conclude that "past research relying on a single self-rating of personality traits has underestimated the true importance of personality for workplace behavioral outcomes."[63] Of course, the trick for real-life personnel selection is gathering that "significant other" data. It appears clear that if you can get such data from peers or others, the aggregated data would probably make for better predictions than self-report measures.

There is no question that personality and other noncognitive attributes are important for understanding and predicting job performance. One very interesting study of franchisees found that the use of personality assessment to select franchisees resulted in the increase in sales royalties from $6,500 per month to $52,000 per franchisee.[64] It is the measurement of the noncognitive attributes in a valid manner that poses challenging problems for HR. It is clear that aggregated observer ratings are strong predictors of future performance. Ratings from multiple (and qualified) peers (and others) can yield predictive validities substantially greater than and incremental to self-report data. Thus the use of both self-report and significant other assessment of personality is recommended.

Let's examine some newer approaches to non-cognitive attribute assessment next.

Approaches to the Prediction of Particular Criteria

There is growing evidence that the use of "compound" traits that are more tied to particular work situations and particular criteria can enhance prediction above what can be derived from the traditional FFM instruments. Many forms of personality, dispositional, or motivation assessment attempt to focus on either particular problems or criteria characteristic of the workplace. Examples are the prediction of voluntary turnover and the prediction of employee theft. One instrument attempts to measure job compatibility in order to predict turnover. Other new instruments are designed to address particular employment issues or situations, such as customer service, violence, or accident proneness.

Predicting (and Reducing) Voluntary Turnover

Employee turnover can be a serious and costly problem for organizations. You may recall the discussion of Domino's Pizza. They found that the cost of turnover was $2,500 each time an hourly employee quit and $20,000 each time a store manager quit. Among other things, Domino's implemented a new and more valid test for selecting managers and hourly personnel that was aimed at predicting both job performance and voluntary turnover. As of 2008, the program was a success on all counts. Turnover was down, store profits were up, and the stock was doing well in an otherwise terrible market. Attracting

Employee referrals reduce turnover

and keeping good employees was a key factor in its turnaround. There are numerous other examples of companies that have expensive and preventable high levels of turnover that can be reduced with better HR policy and practice. Recall the discussion of SAS, the North Carolina software company. Even at the height of the so-called high-tech bubble in the late 1990s, SAS had turnover rates that were well below the industry average. Attracting and keeping good employees is considered a key to the SAS success story. As of 2011, SAS remained one of *Fortune*'s "Best Companies to Work For" and reported its usual very low turnover rate among its core personnel.

One study provided guidelines regarding methods that have been shown to be effective at reducing voluntary turnover.[65] A summary of the findings merged with previous research on turnover is presented in Figure 6-6. This research drew several conclusions. First, voluntary turnover is less likely if a job candidate is referred by a current employee or has friends or family working at the organization. Candidates with more contacts within the organization are apt to better understand the nature of the job and the organization. Such candidates probably have a more realistic view of the job that may provide a "vaccination effect" that lowers expectations, thereby preventing job dissatisfaction and turnover (realistic job previews can also do this). Also, current job holders are less likely to refer job candidates who they feel are less capable or those who (they feel) would not fit in well with the organization's culture.

Another argument for an employee referral system is that having acquaintances within the organization is also likely to strengthen an employee's commitment to the firm and thus reduce the probability that he or she will leave. Of course, this argument also applies to the employee who made the referral.

Another reliable predictor of longer tenure in a job (the opposite of voluntary turnover) is longer tenure in previous jobs. In general, if a person has a history of short-term employment, that person is more likely to quit the next job sooner. This tendency may also reflect a lower work ethic (lower Conscientiousness), which is also correlated with organizational

Figure 6-6 Predictors of Voluntary Turnover and How to Avoid It

1. *Rely on employee referrals*
   - Voluntary turnover is less likely if a job candidate is referred by a current employee or has friends or family working at the organization.
   - Candidates with more contacts within the organization are apt to better understand the nature of the job and the organization.
   - Having friends or family within the organization prior to hire is likely to strengthen the employee's commitment to the firm and reduce the likelihood that he or she will leave.
2. *Put weight on tenure in previous jobs*
   - A past habitual practice of seeking out short-term employment predicts future short-term employment.
   - Short-term employment may reflect a poor work ethic, which is correlated with lack of organizational commitment and turnover.
3. *Measure intent to quit*
   - Intention to quit is one of the best (if not the best) predictors of turnover.
   - Despite their transparency, expressions of intentions to stay or quit before a person starts a new position are an effective predictor of subsequent turnover (e.g., how long do you plan to work for the company?).
4. *Measure the applicant's desires/motivations and job compatibility for the position*
   - New employees with a strong desire for employment will require less time to be assimilated into the organization's culture.
   - Job compatibility is correlated with job tenure.
5. *Use disguised-purpose dispositional measures*
   - Persons with high self-confidence should respond more favorably to the challenges of a new environment.
   - Employees with higher confidence in their abilities are less likely to quit than those who attribute their past performance to luck.
   - Decisive individuals are likely to be more thoughtful about their decisions, more committed to the decisions they make, and less likely to leave the organization.
   - Decisiveness is a component of the personality trait of Conscientiousness from the five-factor model.
   - Decisiveness affects organizational commitment and, indirectly, turnover.
   - High Conscientiousness and high Agreeableness are related to longer tenure.

Sources: Adapted from: Zimmerman, R. D. (2008). Understanding the impact of personality traits on individuals' turnover decisions: A meta-analytic path model. *Personnel Psychology, 61,* 309–348; Barrick, M. R., & Zimmerman, R. D. (2005). Reducing voluntary, avoidable turnover through selection. *Journal of Applied Psychology, 90*, 159–166.

Use WABs to lower turnover for entry-level jobs

commitment and turnover. As discussed earlier, tenure in previous jobs, measured in a systematic manner as a part of a **weighted application blank** (WAB), is predictive of turnover. Intention to quit is also a reliable predictor of, and perhaps the best predictor of, quitting. Believe it or not, questions on an application form such as "How long do you think you'll be working for this company?" are quite predictive of voluntary turnover. Pre-hire dispositions or behavioral intentions, derived from questions such as this one or from interview questions, work quite well.

Measures of the extent of an applicant's desire to work for the organization also predict subsequent turnover. However, almost all of the research on WABs has involved entry-level and nonmanagerial positions, so applicability to managerial positions is questionable. This is not true for biodata (or BIBs).

Disguised-purpose attitudinal scales, where the scoring key is hidden, measuring self-confidence and decisiveness have been shown to predict turnover for higher-level positions as well, including managerial positions. Answers to questions such as "How confident are you that you can do this job well?" or responses to statements like "When I make a decision, I tend to stick to it" also predict turnover quite well. In addition, there is little evidence of adverse impact against protected classes using these measures. This research also revealed that disguised-purpose measures added incremental validity to the prediction of turnover beyond what could be predicted by biodata alone. Personality traits have an impact on individuals' turnover intentions and behaviors. While **Emotional Stability** (from the FFM) is the best predictor (negatively) of employees' intentions to quit, low scores on **Conscientiousness** and **Agreeableness** are the best predictors of actual turnover decisions. Also, individuals who are low on **Agreeableness** or high on **Openness to Experience** may engage in unplanned quitting.

Job Compatibility Questionnaire

Another example of a disguised-purpose dispositional measure is the **Job Compatibility Questionnaire** (JCQ). As discussed in Chapters 4 and 5, the JCQ was developed to determine whether an applicant's preferences for work characteristics matched the actual characteristics of the job.[66] The underlying theory of the JCQ approach is that the compatibility or preference for certain job characteristics will predict job tenure and performance. Test takers are presented groups of items and are instructed to indicate which item is most desirable and which is least desirable. As discussed in Chapter 4, the items are grouped based on a job analysis that identifies those characteristics that are actually descriptive of the job(s) to be filled. Here is an example of a sample group: (a) being able to choose the order of my work tasks, (b) having different and challenging projects, (c) staying physically active on the job, (d) clearly seeing the effects of my hard work. The items are grouped together in such a way that the scoring key is hidden from the respondent, reducing the chance for faking.

Studies involving customer service representatives, security guards, and theater personnel indicate that the JCQ can successfully predict employee turnover for low-skilled jobs. In addition, no evidence of adverse impact has been found. BA&C incorporated the JCQ in its test for security guards. The JCQ has never been used or validated for managerial positions and is not recommended for the selection of managers.

Can We Predict Employee Theft?

It is estimated that employee theft exceeds $400 billion annually. In response to this huge problem and in addition to more detailed background and reference checks, more than 3 million job applicants took some form of honesty or integrity test in 2012. These tests are typically used for jobs in which workers have access to money, such as retail stores, fast-food chains, and banks. Integrity or honesty tests have become more popular since the polygraph, or lie detector, test was banned in 1988 by the **Employee Polygraph Protection Act.** This federal law outlawed the use of the polygraph for selection and greatly restricts the use of the test for other employment situations. There are some employment exemptions to the law, such as those involving security services, businesses involving controlled substances, and government employers.

Integrity/Honesty tests

Integrity/honesty tests are designed to measure attitudes toward theft and may include questions concerning beliefs about how often theft on the job occurs, judgments of the punishments for different degrees of theft, the perceived ease of theft, support for excuses for stealing from an employer, and assessments of one's own honesty. Most inventories also

ask the respondent to report his/her own history of theft and other various counterproductive work behaviors (CWBs).

Sample items typically cover beliefs about the amount of theft that takes place, asking test takers questions such as the following: "What percentage of people take more than $1.00 per week from their employer?" The test also questions punitiveness toward theft: "Should a person be fired if caught stealing $5.00?" The test takers answer questions reflecting their thoughts about stealing: "Have you ever thought about taking company merchandise without actually taking any?" Other honesty tests include items that have been found to correlate with theft: "You freely admit your mistakes." "You like to do things that shock people." "You have had a lot of disagreements with your parents."

Strong validity but few studies predict actual theft

The validity evidence for integrity tests is fairly strong with little adverse impact. Still, critics point to a number of problems with the validity studies. First, most of the validity studies have been conducted by the test publishers themselves; there have been very few independent validation studies. Second, few of the criterion-related validity studies use employee theft as the criterion. A report by the American Psychological Association concluded that the evidence supports the validity of some of the most carefully developed and validated honesty tests. The most recent studies on integrity tests support their use.[67] Although designed to predict CWBs, especially employee theft, integrity tests have also been found to predict job performance in general. One major study found that integrity tests had the highest incremental validity (of all other tests) in the prediction of job performance beyond GMA.[68] Scores on integrity tests are also related to Conscientiousness, Emotional Stability, and Agreeableness of the FFM. It has been proposed that a trait represented on integrity tests is not well represented by the FFM. "Honesty-Humility (H-H)" has been proposed as the sixth factor defined as "sincerity, fairness, lack of conceit, and lack of greed." There is evidence that this sixth factor can enhance the prediction of CWBs or workplace delinquency.[69]

Highest incremental validity with GMA

Can We Identify Applicants Who Will Provide Good Customer Service?

Considerable research demonstrates that employees' customer orientation is a good predictor of customer-related outcomes such as customer and supervisory ratings of service performance, customer-focused organizational citizenship behaviors, and customer satisfaction. Thus, identifying employees who would have such an orientation would be advantageous for organizations with a strong customer-focused strategy. The **Service Orientation Index (SOI)** was initially developed as a means of predicting the helpfulness of nurses' aides in large, inner-city hospitals.[70] The test items were selected from three main dimensions: patient service, assisting other personnel, and communication. Here are some examples of SOI items: "I always notice when people are upset" and "I never resent it when I don't get my way." Several other studies of the SOI involving clerical employees and truck drivers have reported positive results as well.

Can We Identify Bad and Risky (and Costly) Drivers?

Driving accidents by employees can be a very costly expense for employers where driving to and from jobs is an essential function of the job. Think cable companies, UPS, FedEx, and exterminators for a few examples of companies that should pay careful attention to the "accident proneness" of the drivers they hire. In addition, employers are often held responsible for the driving behavior of their employees when they are on the job. A plethora of **negligent hiring** lawsuits have looked at what screening procedures were used to hire the guy who committed a driving infraction while on the job and caused a serious accident.

So, first off, is there such a thing as "accident proneness," and if so, can we predict it in job applicants? The answers to these two key questions are in fact "yes" and "yes." Research shows that a person's previous driving record is the single best predictor of the on-the-job record and an essential screening tool. But personality is a correlate of risky driving behavior and future traffic violations and accidents. For young drivers (18–25), one study found that a high level of "thrill-seeking" and aggression, combined with a low level of empathy, was a predictor of subsequent risky driving and speeding violations. The researchers measured these subfactors from the "Big-Five" traits. The subfactors derived from the Emotional Stability (anger/aggression), Extraversion ("thrill-seeking"), and Agreeableness (low empathy) components of the FFM.[71]

Driving record is a strong predictor

Personality predicts risky driving

Accident-proneness can be predicted

Another test developed to predict (and prevent) accidents is the **Safety Locus of Control Scale** (SLC), which is a paper-and-pencil test containing 17 items assessing attitudes toward safety. A sample item is as follows: "Avoiding accidents is a matter of luck." Validity data look encouraging across different industries, including transportation, hotels, and aviation. In addition, these investigations indicate no adverse impact against minorities and women.[72]

Results with older drivers also suggest that a "sensation-seeking" personality and low levels of emotional stability are related to risky driving among older drivers in addition to cognitive and motor abilities.[73] The perception of reckless driving as acceptable and desirable or as negative and threatening and the risk assessment related to cell phone usage are other predictors of driving behavior and accidents. There apparently is such a thing as "accident prone" in the sense that the people most "prone" to be involved in accidents can be identified with a background check and a personality inventory.

How Do You Establish a Testing Program?

Establishing a psychological testing program is a difficult undertaking—one that should ideally involve the advice of an industrial psychologist. HR professionals should follow these guidelines before using psychological tests.

1. Most reputable testing publishers provide a test manual. Study the manual carefully, particularly the adverse impact and validity evidence. Has the test been shown to predict success in jobs similar to the jobs you're trying to fill? Have adverse impact studies been performed? What are the findings? Are there positive, independent research studies in scholarly journals? Have qualified experts with advanced degrees in psychology or related fields been involved in the research?
2. Check to see if the test has been reviewed in *Mental Measurements Yearbook (MMY).* Published by the Buros Institute of the University of Nebraska, the MMY publishes scholarly reviews of tests by qualified academics who have no vested interest in the tests they are reviewing. You can also download Buros test reviews online at http://buros.unl.edu/buros/jsp/search.jsp. You can retrieve reviews by test name or by category (e.g., achievement, intelligence, personality).
3. Ask the test publishers for the names of several companies that have used the test. Call a sample of them and determine if they have conducted any adverse impact and validity studies. Determine if legal actions have been taken related to the test; if so, what are the implications for your situation?
4. Obtain a copy of the test from the publisher and carefully examine all of the test items. Consider each item in the context of ethical, legal, and privacy ramifications. Organizations have lost court cases because of specific items on a test.

Proceed cautiously in the selection and adoption of psychological tests. Don't be wowed by a slick test brochure; take a step back and evaluate the product in the same manner you would evaluate any product before buying it. Be particularly critical of vendors' claims and remember that you can assess personality and motivation using an interview. If you decide to adopt a test, maintain the data so that you can evaluate whether the test is working. In general, it is always advisable to contact someone who can give you an objective, expert appraisal.

DRUG TESTING

Drug abuse is one of the most serious problems in the United States today with productivity costs in the billions of dollars and on the rise. Drug abuse in the workplace also has been linked to employee theft, accidents, absences, use of sick time, and other counterproductive behavior.

Many organizations are turning to drug testing for job applicants and incumbents. One survey found that 87 percent of major U.S. corporations now use some form of drug

testing.[74] While some of the tests are in the form of paper-and-pencil examinations, the vast majority of tests conducted are clinical tests of urine or hair samples. Ninety-six percent of firms refuse to hire applicants who test positive for illegal drug use, methamphetamines, and some prescription drugs (e.g., OxyContin). While the most common practice is to test job applicants, drug testing of job incumbents, either through a randomized procedure or based on probable cause, is also on the increase.

The most common form of urinalysis testing is the immunoassay test, which applies an enzyme solution to a urine sample and measures change in the density of the sample. The drawback of the $20 (per applicant) immunoassay test is that it is sensitive to some legal drugs as well as illegal drugs. Because of this, it is recommended that a positive immunoassay test be followed by a more reliable confirmatory test, such as gas chromatography. The only errors in testing that can occur with the confirmatory tests are due to two causes: positive results from passive inhalation, a rare event (caused by involuntarily inhaling marijuana), and laboratory blunders (e.g., mixing urine samples). Hair analysis is a more expensive but also more reliable and less invasive form of drug testing. Testing for methamphetamine use is difficult since the ingredients pass through the body quickly.

Positive test results say little regarding one's ability to perform the job, and most testing gives little or no information about the amount of the drug that was used, when it was used, how frequently it was used, and whether the applicant or candidate will be (or is) less effective on the job.

Drug testing is legal in all 50 states

The legal implications of drug testing are evolving. Currently, drug testing is legal in all 50 states for preemployment screening and on-the-job assessment; however, employees in some states have successfully challenged dismissals based solely on a random drug test. For those employment situations in which a collective-bargaining agreement has allowed drug testing, the punitive action based on the results is subject to arbitration. One study found that the majority of dismissals based on drug tests were overturned by arbitrators.[75] Among the arguments against drug testing are that it is an invasion of privacy, it is an unreasonable search and seizure, and it violates the right of due process. Most experts agree that all three of these arguments may apply to public employers, such as governments, but do not apply to private industry. State law is relevant here since some drug testing programs have been challenged under privacy provisions of state constitutions. With regard to public employment, the Supreme Court has ruled that drug testing is legal if the employer can show a "special need" (e.g., public safety).[76] Drug testing is covered in more detail in Chapter 14.

Is Some Testing an Invasion of Privacy?

Politically-oriented questions are illegal in some states

The widespread use of various employment tests has been criticized on the grounds that these procedures may be an invasion of individuals' privacy and unnecessarily reveal information that will affect individuals' employment opportunities. Selection methods that seem to provoke these concerns are drug tests, personality tests, and honesty/integrity tests. Questions on tests or interviews that are political in tone are illegal in some states. Experts in the field of employment testing who support testing have responded to this challenge in a number of ways. First, various professional standards and guidelines have been devised to protect the confidentiality of test results. Second, since almost any interpersonal interaction, whether it be an interview or an informal discussion with an employer over lunch, involves the exchange of information, advocates of employment testing contend that every selection procedure compromises applicants' privacy to some degree. Finally, in the interests of high productivity, and staying within the law, they assert, organizations may need to violate individuals' privacy to a certain extent. Companies with government contracts are among those that are obliged to maintain a safe work environment and may need to require drug testing and extensive background checks of employees.

Concerns will continue to be voiced over the confidentiality and ethics of employment testing, particularly as computer-based databases expand in scope and availability to organizations. It is also likely that there will be increasing calls for more legislation at federal, state, and local levels to restrict company access to and use of employment-related information.

PERFORMANCE TESTING/ WORK SAMPLES

Despite making valuable contributions to employee selection, GMA tests have their problems and limitations. The validity of GMA is proven and clear. Unfortunately, the potential legal implications of their use persist. However, the validity of self-report measures of motivation or personality is not nearly as impressive. Many experts suggest that the prediction of job performance can be enhanced through **performance testing,** which is the sampling of simulated job tasks and/or behaviors. There is also evidence that the use of such tests can result in less adverse impact than GMA tests and that test takers perceive such tests as more accurate and fair.[77] However, while such tests may reduce AI compared to the exclusive use of GMA or knowledge-based tests, recent evidence indicates that differences are larger than previously thought and that differences are larger when the underlying constructs being assessed are knowledge or cognitive ability while differences are smaller when the underlying constructs concern various social skills.[78]

Performance tests and work samples have good validity

Performance tests measure KASOCs or competencies (e.g., application of knowledge or a skill in a simulated setting). Performance tests involve actual "doing" rather than "knowing how." Thus, a performance test may require a job candidate to demonstrate a skill such as written communication or analytical ability. Applicants may also be required to prepare something for a live demonstration. Thus, preparing a lesson plan for a unit of instruction could be the first step before a simulated class is conducted.

Work sample tests are exercises that reflect actual job responsibilities and tasks. Applicants are placed in a job situation and are required to handle tasks, activities, or problems that match those found on the job. The purpose of a simulation or work sample test is to allow applicants to demonstrate their job-related competencies in as realistic a situation as possible.

Work samples can duplicate a real-life event but eliminate the risks of danger or damage such as substituting safe substances or chemicals to test the correct handling of dangerous materials or using driving or flight simulators. Like performance tests, work samples are conducted under controlled conditions for the purposes of consistency and fairness and can be developed using a number of different formats. To ensure that performance tests and work samples are tailored to match the important activities of the job, HR professionals should develop the methods from the tasks, behaviors, and responsibilities identified in a job analysis (see Chapter 4).

One example of a sophisticated approach to work samples (and recruiting) is Google's "Code Jam," an international programming competition administered by Google. Introduced in 2003, Google uses Code Jam results to identify top engineering talent for potential employment at Google. The one-day competition consists of a set of complex programing problems that must be solved in a fixed (and short) amount of time. For example, competitors have been asked to develop a complex war game in less than 2 hours. Google has had great success with this approach to recruiting and hiring. It claims that it has been able to hire over 50 percent of the finalists every year since 2003.[79]

Another form of performance testing is the **Situational Judgment Test (SJT).** This test consists of a number of job-related situations presented in written, verbal, or visual (video) form. Unlike a typical work sample, SJTs present hypothetical situations and ask respondents how they would respond. Here's an example of an SJT question.[80]

> A customer asks for a specific brand of merchandise the store doesn't carry. How would you respond?
> A. Tell the customer which stores carry that brand, but point out that your brand is similar.
> B. Ask the customer more questions so you can suggest something else.
> C. Tell the customer that the store carries the best merchandise available.
> D. Ask another associate to help.
> E. Tell the customer which stores carry the brand.
> Questions:
> **1.** Which of the options above do you believe is the best under the circumstances?
> **2.** Which of the options above do you believe is the worst under the circumstances?

SJTs have incremental validity

Research on SJTs is quite positive.[81] SJTs most often assess leadership and interpersonal skills and have relatively high validities for predicting overall job performance. In addition, video-based situational judgment tests have stronger criterion-related validity than pencil-and-paper measures.[82] SJTs have incremental validity above GMA, personality, and job/training experiences measures. The SJT approach has also been shown to be a promising and valid predictor of personal initiative.[83] Recent research indicates that scoring keys derived by subject matter experts and items saturated with specific knowledge about effective job behavior result in higher validities compared to other approaches to deriving scoring keys for SJTs.[84]

The performance testing process should be standardized as much as possible with consistent and precise instructions, testing material, conditions, and equipment. All of the candidates must have the same time allotment to complete tests, and there must be a specific standard of performance by which to compare the applicants' efforts. To illustrate the point, a minimum passing score for a typing exam might be set at 40 words a minute with two errors. This standard would apply to all the applicants. Today, performance tests are available through the Internet. One large retailer had candidates for its district manager position complete a performance test over a website. Once responses are made through the website, trained assessors conduct interviews that focus on the candidates' responses.

Web-based testing

Although the research is limited, that which exists tends to support proctored, web-based testing.[85] Studies involving SJTs, biodata, and personality measurement using the Five-Factor Model indicate that proctored, web-based testing has positive benefits relative to paper-and-pencil measures. Research shows that validity coefficients can exceed .60 with a combination of work-sample tests, a structured interview, and a measure of GMA.[86]

What Is an Assessment Center?

An **assessment center** is a collection of many of the selection tools already discussed. The use of multiple techniques and a standardized process of data collection contributes to the validity of the method. Unlike most of the research on GMA, most of the validity evidence on assessment centers is from studies of management positions. These "centers" use trained observers and a variety of techniques to make judgments about behavior, in part, from specially developed assessment simulations. **Assessors typically test job candidates with a collection of performance tests that simulate the work environment.** Some centers also use paper-and-pencil tests, including GMA and personality tests, as part of the assessment process. At the Center for Creative Leadership in Greensboro, North Carolina, managers complete a battery of cognitive and personality tests and receive subordinate and peer assessments prior to their participation in the 2-day assessment center, which includes five performance tests.

Private sector organizations, educational institutions, military organizations, public safety, and other governmental agencies have used the assessment center method to identify candidates for selection, placement, and promotion. Because of the cost, most organizations restrict use of assessment centers to only supervisory and managerial selection. There have been some applications of the method for nonadministrative positions such as sales personnel, vocational rehabilitation counselors, planning analysts, social workers, personnel specialists, research analysts, firefighters, and police officers.

Allows for direct comparisons among internal and external candidates

One of the advantages of the assessment center approach for managerial selection is that internal and external candidates can go through the assessment center to provide a direct comparison of the candidates, as they participate (and compete) in the collection of performance tests. Candidates are assessed and compared by trained assessors. Among the numerous organizations that use the assessment center method for selection are the FBI, AT&T, IBM, Ford, Office Depot, Xerox, Procter and Gamble, the Department of Defense, the CIA, and the Federal Aviation Administration. Assessment centers are expensive with costs ranging from a low of about $300 for each candidate to as much as $8,000 for upper-level managerial selection.

Features trained assessors and multiple performance tests

With the typical assessment center method, information about an employee's strengths and weaknesses is provided through a combination of performance tests that are designed to simulate the type of work to which the candidate will be exposed. A team of trained assessors observes and evaluates performance in the simulations. The assessors compile and

integrate their judgments on each exercise to form a summary rating for each candidate being assessed. Assessment centers tend to vary in terms of length of the assessment process (1 day to 1 week), the ratio of assessors to those being assessed, the extent of assessor training, and the number and type of assessment instruments and exercises that are used to assess candidates.[87]

Assess job dimensions or competencies

All assessment centers call for an assessment of job *dimensions* or competencies. For example, United Technology evaluates managers on the following dimensions: oral presentation, initiative, leadership, planning and organization, written communication, decision making, and interpersonal skills. These **dimensions or competencies are clusters of behaviors that are specific, observable, and verifiable and can be reliably and logically classified together.** The dimension "written communication" was defined by United Technology as the following: "clear expression of ideas in writing and in good grammatical form." United Technology breaks down behavioral examples of written communication as: "Exchanges information/reports with superior regarding the day's activities. Completes all written reports and required forms in a manner that ensures the inclusion of all data necessary to meet the needs of the personnel using the information. Uses appropriate vocabulary and avoids excessive technical jargon in required correspondence." Figure 6-7 presents a set of dimensions and their definitions as used in an assessment center for selecting supervisors. There are essentially no differences between "competencies" and dimensions as they are typically defined.

The assessment dimensions, performance tests or exercises are developed from the results of a job analysis. The exercises allow assessors to observe, record, classify, and evaluate relevant job behaviors. Some of the most common assessment exercises are **in-baskets, leaderless group discussions, oral presentations, and role-playing.** Descriptions of these methods follow.

In-Basket

The *in-basket* consists of a variety of materials of varying importance and priority that typically would be handled by a manager the organization is trying to hire. Candidates are asked to imagine that they are placed in the position and must deal with a number of memos and items accumulated in their in-baskets. Assessors give them background information about the unit they are managing, and they must deal with the in-basket materials in a limited amount of time. After writing their responses to the memos, the candidates are interviewed by trained assessors who review the "out-basket" and question the actions

Figure 6-7 Assessment Center Dimensions: An Example

Leadership: To direct, coordinate, and guide the activities of others; to monitor, instruct, and motivate others in the performance of their tasks; to assign duties and responsibilities and to follow up on assignments; to utilize available human and technical resources in accomplishing tasks and in achieving solutions to problems; to follow through within organizational guidelines.

Interpersonal: To be sensitive to the needs and feelings of others; to respond empathetically; to consistently display courtesy in interpersonal contacts; to develop rapport with others; to be cognizant of and respect the need in others for self-esteem.

Organizing and Planning: To create strategies for self and others to accomplish specific results; to utilize prescribed strategies; to fix schedules and priorities so as to meet objectivities; to coordinate personnel and other resources; to establish and utilize follow-up procedures.

Perception and Analysis: To identify, assimilate, and comprehend the critical elements of a situation; to identify alternative courses of action; to be aware of situational or data discrepancies; to evaluate salient factors and elements essential to resolution of problems.

Decision Making: To use logical and sound judgment in use of resources; to adequately assess a situation and make a sound and logical determination of an appropriate course of action based on the facts available, including established procedures and guidelines; to select solutions to problems by weighing the ramifications of alternative courses of action.

Oral and Nonverbal Communication: To present information to others concisely and without ambiguity; to articulate clearly; to use appropriate voice inflection, grammar, and vocabulary; to maintain appropriate eye contact; to display congruent nonverbal behavior.

Adaptability: To modify courses of action to accommodate situational changes; to vary behavior in accordance with changes in human and interpersonal factors; to withstand stress.

Decisiveness: To make frequent decisions; to make decisions spanning many different areas; to render judgments, take action, and make commitments; to react quickly to situational changes; to make determinations based on available evidence; to defend actions when challenged by others.

Written Communications: To present and express information in writing, employing unambiguous, concise, and effective language. To use correct grammar, punctuation, and sentence structure; to adjust writing style to the demands of the communication.

taken. In-baskets are typically designed to measure oral and written communication skills, planning, decisiveness, initiative, and organization skills.

Leaderless Group Discussion

Candidates assemble in groups of three to six people after individually considering an issue or problem and making specific recommendations. While a leader is not designated for the group, one usually emerges in the course of the group interaction. Two or more assessors observe the interaction as the group attempts to reach consensus on the issue. Assessors typically use the leaderless group discussion to determine oral communication, stress tolerance, adaptability, leadership, and persuasiveness. Some graduate schools now use the leaderless group discussion to select doctoral students for their business and other graduate programs.

Oral Presentation

In the brief time allowed, candidates plan, organize, and prepare a presentation on an assigned topic. An assessment center developed by IBM requires candidates for sales management positions to prepare and deliver a 5-minute oral presentation in which they present one of their hypothetical staff members for promotion, and then defend the staff member in a group discussion. IBM uses this exercise to evaluate assertiveness, selling ability, self-confidence, resistance, and interpersonal contact.[88]

Role-Playing

For this common assessment center exercise, candidates assume the role of the incumbent and must deal with a subordinate about a performance problem. The subordinate is a trained role-player. Another example is to have candidates interact with clients or individuals external to the organization, requiring them to obtain information or alleviate a problem. Vocational rehabilitation counselor candidates who apply for jobs with the Massachusetts Rehabilitation Commission assume the role of a counselor who is meeting a client for the first time. The candidate has the responsibility of gaining information on the client's case and establishing rapport with the client. Figure 6-8 presents summary descriptions of four exercises used in an assessment center to select store managers in a retail environment.

How Are Assessments Done?

Assessors who have received extensive training on assessment center methodology evaluate all of the candidates in an assessment center—usually 6 to 12 people—as they perform the same tasks. Assessors are trained to recognize designated behaviors, which are clearly defined prior to each assessment.

Assessors are often representatives from the organization who are at higher levels than the candidates being assessed. This is done to diminish the potential for contamination,

Figure 6-8 **Description of Assessment Center Exercises for Retail Managers**

Customer Situation: A large equipment user (a select national account) has been experiencing recent problems involving a particular piece of equipment, culminating in a systems-down situation. Problems with the equipment could include software, and parts received to fix the equipment are damaged.

The participant will be required to review information about the problem for 30 minutes and generate potential courses of action. Participants will then meet in groups to devise a consensus strategy for dealing with the problem. Assessors should expect a plan of action from the participants and may probe the participants for additional contingency plans. The participants will have 45 minutes to discuss the customer problem and develop a strategy.

Employee Discussion: In this exercise the participant must develop a strategy for counseling a subordinate (a senior customer service engineer) who has been experiencing recent performance problems. The participant will have 30 minutes to review information regarding the technician's declining performance over the last few months.

The participant will then have 15 minutes to prepare a brief report on the individual with recommendations for submission to the district manager. The participant will then meet with two assessors to discuss the strategy.

In-Basket: In this exercise, the participant will assume the role of a newly transferred branch manager. The participant will have 90 minutes to review information related to various issues (technical developments, equipment maintenance specifications, customer information, etc.). The participant will be instructed to spend this time identifying priorities and grouping related issues, as well as indicating courses of action to be taken. The participant will then take part in a 15-minute interview with an assessor to clarify the actions taken and logic behind decisions made.

Problem Analysis: In this exercise the participant will be required to review information on three candidates and provide a recommendation on which of the three should be promoted to a branch manager position. The participant will have 90 minutes to review information and prepare a written recommendation. The participants will then meet in groups to derive a consensus recommendation for the district manager.

which may result from an assessor allowing prior association with a candidate to interfere with making an objective evaluation. Some assessment centers use outside consultants and psychologists as assessors and there is some evidence that this will increase validity.

Different assessors observe assessment center candidates in each exercise. The assessors are responsible for observing the actual behavior of the candidate during each exercise and documenting how each candidate performed.

After the participants complete all of the exercises, the assessors typically assemble at a team meeting to pool their impressions, arrive at an overall consensus rating for each candidate on each dimension, and derive an overall assessment rating.

There is some evidence that assessment centers can be broken down to make them less costly and more efficient. Research shows that you probably do not have to assemble candidates together at a "center"; performance tests completed online and follow-up interviews by trained assessors reveal essentially the same results as the more typical assessment centers.[89]

What Is the Validity and Adverse Impact of Assessment Centers and Other Performance Tests?

There is a scarcity of well-done, criterion-related validity studies on assessment centers. With a few exceptions, assessment center validity studies focus on administrative positions such as managers and supervisors.[90] The method also has proved to be valid for law enforcement personnel.[91] In general, **the validity of assessment centers is strong**[92] (see Figure 6-2), particularly for managerial positions. Also, research indicates that higher criterion-related validity can be obtained when fewer dimensions are used and when assessors are psychologists.[93]

Strong validity for managerial positions

While the validities reported for assessment centers are similar to those reported for GMA, decisions made from assessment centers are more defensible in court and result in less adverse impact than cognitive ability tests.[94] **The method is ideal when an organization has both internal and external candidates.** Most companies use assessment centers as one of the last steps in a selection process where a limited number of internal and external candidates are being considered. People who are assessed by the assessment center method or performance tests perceive the procedure to be fair and job related, making them less likely to take legal action.

More defensible in court (less AI than GMA)

Performance Appraisals/ Competency Assessment

The use of competencies as a fundamental building block of organizations and the people they employ is increasingly popular and is often used as the basis for personnel decisions within an organization. Remember that a policy of promotion from within the organization (based to some extent on past performance in other jobs) is a **High-Performance Work Practice related to subsequent corporate performance.** But there is little research on the validity of performance-based competency assessment or performance appraisal in general for predicting performance at a higher level. Does high performance in Job A, for example (at least as rated by supervisors, co-workers, or others), predict performance in Job B? Many organizations use promotability ratings although there has been little research on these judgments. Most of these judgments come from an employee's immediate supervisor. One study examined the relationship between employees' challenging job experiences and supervisors' evaluations of employees' promotability over and above the employees' current job performance. Results showed that challenging job experiences explained differences in evaluations of promotability over and above current job performance and job tenure.[95] Based on related research, it is clear that judgments of promotability should come from more than one source and that, if possible, several peers should be used in this process.[96]

Many organizations now use some form of a multirater or 360-degree assessment process to measure competencies. Appraisal data can often be found in human resource information systems (HRIS) and used for succession planning. PeopleSoft's most popular HRIS, for example, includes a web-based competency-appraisal system, the data of which are maintained on each employee and help companies do succession and career planning.

But how does 360-degree appraisal or, for that matter, appraisal from any rating source compare on its ability to predict later performance relative to some of these other tools just described? Is 360-degree appraisal data, or peer assessment, or supervisory assessment as

good as (or better than) assessment centers or testing, for example? One study in a retail environment addressed this issue comparing the levels of criterion-related validity and the extent of statistical adverse impact against minorities with three popular methods.[97] Data based on top-down (supervisory) performance appraisals, a 360-degree competency-based appraisal system, and a traditional assessment center were correlated with subsequent job performance of retail store managers. The assessment center and 360-degree systems had the highest levels of predictive validity while the "top-down" managerial assessment was significantly lower (.46 for ACs, .37 for 360-degree versus .19 for "top-down"). The 360-degree data and the assessment center (AC) also resulted in less adverse impact than the "top-down" method.

360-degree PA had higher validity than "top-down" appraisal

Evidence for the incremental validity of 360-degree appraisal data above the AC data was also found, indicating more accurate prediction with the combination of AC and 360-degree data. While this one study showed practical usefulness for the 360-degree appraisal as a source of data for personnel decisions, these data are obviously problematic if both internal and external candidates are being considered, since no 360-degree data would be available for the external candidates. However, you should not ignore useful (and valid) information because some candidates do not have it. Use whatever *valid* data you have but, if possible, try to obtain the full complement of data on all candidates. This is one advantage of assessment centers for higher-level staffing decisions. When you have external candidates competing against internal candidates for managerial positions, assessment centers create a "level playing field" of valid sources of information about the candidates.

Incremental validity for 360 appraisal with AC data

INTERVIEWS

While the use of paper-and-pencil tests and performance tests has increased, the employment interview continues to be the most common personnel selection tool. Primarily due to its expense, the interview is typically one of the last selection hurdles used after other methods have reduced the number of potential candidates. The manner in which interviews are conducted is not typically conducive to high validity for the method. But there is clear evidence that interviews, when done properly, can be quite valid.

One of the bigger discrepancies between HRM research and practice is in the area of interviewing. Research provides clear prescriptions for interviewing the right way and this way is clearly at odds with the way it is typically done. Figure 6-9 presents the most important discrepancies between research and practice as related to interviewing based on a survey of 164 HR managers working for organizations with 100 or more employees.

Figure 6-9
Discrepancies between Research and Practice for Employment Interviews

What Does Research Say?	What Is the Practice?
Use job analysis to derive questions	22% of companies use formal job analysis
Monitor interview data for adverse impact	26% of companies do
Validate interview format/content	20% of companies do
Train interviewers	36% do
Formally weight hiring factors based on job analysis	6% do
Use a structured interview format	17% do
Use "situational" interview questions	34% do
Use "behavioral" interview questions	25% do
Use a formal interview rating system	24% do
Use more than one interviewer	52% do
Use statistical model to combine data from other sources (tests, bio-data, etc.)	2% use actuarial or statistical model

Source: H. J. Bernardin, "The Frequency of Use and Perceived Validity of Staffing Method Options," 2011. Unpublished manuscript.

The good news is that the results reported in Figure 6-9 are an improvement on previous survey results. Even academic institutions, from which the vast majority of this research is derived, do not usually practice what they preach when it comes to selecting a new faculty member or administrator.

Almost every student eventually will take part in a job interview. Nearly l00 percent of organizations use the employment interview as one basis for personnel selection. Even some universities now use interviews to select students for graduate programs. Dartmouth, Carnegie-Mellon, and The Wharton School at the University of Pennsylvania routinely interview applicants for their prestigious MBA programs. Many companies now provide extensive training programs and specific guidelines for interviewers. As Tom Newman, director of training at S. C. Johnson & Son, Inc., said, interviewing is now "much more of a science." This "science" clearly pays off as research shows greater validity for more systematic interviewing. Mobil Oil, Radisson Hotels International, the Marriott Corporation, and Sun Bank are among the many companies with extensive programs to prepare their interviewers.

What Factors Affect the Employment Interview?

A veritable plethora of research has been devoted to the employment interview.[98] This research has focused on the attributes of the applicant, the attributes of the interviewer, extraneous variables that affect interview results, interview formats, and, of course, the validity of interviews related to all of these things.

In the context of the interview, the attributes of the applicant refer to characteristics that influence an interviewer's attention to and impression of the applicant. Voice modulation, body language, posture, interviewee anxiety, and visible characteristics such as sex, weight, ethnicity, and physical attractiveness are among the factors that might influence the interviewer's judgments about a job applicant. A common phenomenon here is "stereotyping," in which an impression about an individual is formed due to his/her group membership rather than any individual attributes. **Stereotyping** involves categorizing groups according to general traits and then attributing those traits to a particular individual once the group membership is known. Although stereotypes are a common and convenient means of efficiently processing information, they can be a source of bias when people attribute traits they believe to be true for an entire group to one member—without considering that person as an individual. Expert witnesses in EEO litigation often cite "stereotyping" as an error more likely to occur when the selection process is **"excessively subjective"** such as an informal, unstructured interview conducted by a single white male.

The interviewer's personal characteristics also can influence his/her judgment in other ways, resulting in interviews that can be characterized as "excessively subjective." Personal values and previously learned associations between certain information cues and decision responses might influence an interviewer's decision-making process. One type of subjective perceptual influence is a "similar-to-me" attribution, meaning the interviewer forms an impression of perceived similarity between an applicant and himself/herself based on the interviewer's attitudes, interests, or group membership, causing certain information, or individuals, to be placed in a more favorable light than others. The danger is that these judgments on the basis of similarity can cause rating errors and bias; the perceived advantages might not be relevant to the particular job for which the interview is being conducted.

"Similar-to-me"

Factors such as stress, background noise, interruptions, time pressures, decision accountability, and other conditions surrounding the interview also can influence interviewers' attention to information. An important factor is the amount of information about the job the interviewer has prior to the actual interview session. Little background information about the job may cause distortion in the decision-making process because of resulting irrelevant or erroneous assumptions about job requirements. This lack of job information causes the interviewer to rely on his/her assumptions about what the job requires. These can be inconsistent across different interviewers or across different interview sessions. Rating errors occur because interviewers collect non-job-related information and use the information to make decisions.

Factors related to attention to information

Thus applicant, interviewer, and situation attributes can potentially bias the decision-making process and result in erroneous evaluations during the interview. In response

Applicant, interviewer, and situational attributes can bias interviews

to these problems, as well as the high cost of face-to-face interviews, many companies conduct computer interviews to screen applicants. Many stores now have a computer workstation where you can complete job application online and take an employment test. Telecomputing Interviewing Services in San Francisco lists more than 1,500 clients that conduct computer interviews for mostly entry-level jobs. Bloomingdale's hires almost all of its entry-level personnel for its Florida stores using computer interviewing that questions applicants about work attitudes, substance abuse, and employee theft. As Ellen Pollin, personnel manager at Bloomingdale's, puts it, "The machine never forgets to ask a question and asks each question in the same way." Many other companies are now using videoconferencing to interview employees, particularly managerial prospects. Texas Instruments claims considerable cost savings with no loss in validity using videoconferences.

Structured and standardized interviewing is growing in popularity. Perhaps the biggest company in this business is the **Gallup Organization** (visit www.gallup.com and find "talent-based hiring" for a description). Gallup conducted a huge study of management behavior, described in the best seller *Now, Discover Your Strengths*.[99] Gallup associates conducted over 1.7 million interviews at 101 companies from 63 countries. One result of this research was a structured interview that is administered by telephone and then scored based on the taped transcript using a standardized rating form. This talent assessment tool is now used by, among many others, Disney, Toyota, Marriott, and Best Buy to help select managers and sales personnel. This nontraditional way to conduct an interview nonetheless resulted in the same level of validity as the more traditional approach.[100]

What Is the Validity of Employment Interviews?

The information obtained from the interview provides a basis for subsequent selection and placement decisions whose overall quality depends on the interview. How reliable is the interview information? How valid is that information for predictive purposes? That is, to what extent do interview judgments predict subsequent job performance and other important criteria?

The validity of the employment interview often has been impaired by underlying perceptual bias owing to factors such as first impressions, stereotypes, different information utilization, different questioning content, and lack of interviewer knowledge regarding the requirements of the job to be filled. However, as a result of recent efforts to improve interview effectiveness, research indicates that certain types of interviews are more reliable and valid than the typical, unstructured format. For instance, interview questions based on a job analysis (see Chapter 4), as opposed to psychological or trait information, increase the validity of the interview procedure.[101] **Structured interviews,** which represent a standardized approach to systematically collecting and rating applicant information, have yielded higher reliability and validity results than unstructured interviews (.43 versus .31). Research findings also suggest that the effectiveness of interview decisions can be improved by carefully defining what information is to be evaluated, by systematically evaluating that information using consistent rating standards, and by focusing the interview (and interview questions) on past behaviors and accomplishments in job-related situations.[102]

Structured interviews have strong validity

There perhaps is a way to high validity, however, without the benefit (and cost) of structured, behavioral interviews based on a thorough job analysis. One study showed that averaging across three or four independent, unstructured interviews is equivalent in validity to a structured interview done by one interviewer.[103]

With potential bias affecting employment interviews comes potential litigation. Many cases have involved the questions that are asked at the interviews. The employment interview is in essence a "test" and is thus subject to the same laws and guidelines prohibiting discrimination on the basis of age, race, sex, religion, national origin, or disability. Furthermore, the interview process is similar to the subjective nature of the performance appraisal process; hence, many of the court decisions concerning the use of performance appraisals also apply to the interview. Judges have not been kind to employers using vague, inadequate hiring standards, "excessive subjectivity," idiosyncratic interview evaluation criteria, or biased questions unrelated to the job. The courts also have criticized employers for inadequate interviewer training and irrelevant interview questions. In general, the courts

Legally, the interview is a test

have focused on two basic issues for determining interview discrimination: the content of the interview and the impact of those decisions.[104]

The first issue involves **discriminatory intent:** Do certain questions convey an impression of underlying discriminatory attitudes? Discrimination is most likely to occur when interviewers ask non-job-related questions of only one protected group of job candidates and not of others. Women applying for work as truck drivers at Spokane Concrete Products were questioned about child care options and other issues not asked of male applicants. The court found disparate treatment against females and a violation of Title VII. An interviewer extensively questioned a female applicant of a bank about what she would do if her 6-year-old got sick. The same interviewer did not ask that question of the male applicants. The applicant didn't get the job but did get a lawyer. The court concluded that this line of questioning constitutes sex discrimination.

Discriminatory impact

The second issue pertains to *discriminatory impact*: Does the interview inquiry result in a differential, or adverse, impact on protected groups? If so, are the interview questions valid and job related? Discriminatory impact occurs when the questions asked of all job candidates implicitly screen out a majority of protected group members. Questions about arrests can have a discriminating impact on minorities. The Detroit Edison Company provided no training, job analysis information, or specific questions for its all-white staff of interviewers. The process could not be defended in light of the adverse impact that resulted from interview decisions.

Watson v. Ft. Worth Bank

Take note that the Supreme Court ruled in *Watson v. Ft. Worth Bank* that "disparate impact" theory may be used for evaluating employment interviews that are used for decision making. An informal, unstructured, and therefore "excessively subjective" interview conducted by "stereotyping" white males will be difficult to defend in the context of evidence of adverse impact in the decisions.

In summary, the inherent bias in the interview and the relatively poor validity reported for unstructured interview decisions make this selection tool vulnerable to charges of both intentional "treatment" and "impact" discrimination. Employers need to quantify, standardize, and document interview judgments. Furthermore, employers should train interviewers, continuously evaluate the reliability and validity of interview decisions, and monitor interviewer decisions for any discriminatory effects. Many companies such as S. C. Johnson, Radisson Hotels, and ExxonMobil now have extensive training programs for interviewers. This training covers interviewing procedures, potential discriminatory areas, rating procedures, and role-plays.

Sex Discrimination

Although early research studies indicated that female applicants generally receive lower interview evaluations than do male applicants, more detailed analyses suggest that this effect is largely dependent on the type of job in question, the amount of job information available to the interviewer, and the qualifications of the candidate. In fact, recent research suggests that females typically do not receive lower ratings in the selection interview; in some studies, females scored higher ratings than male applicants. Of course, this research can be (and has been) used in litigation against an organization where there is evidence of disparate impact against women based on interview decisions.

Race Discrimination

There is mixed evidence for racial bias in interviewer evaluations. Positive and negative results have been reported in the relatively few studies that have investigated race discrimination. There is some indication that African American interviewers rate African American applicants more favorably while white interviewers did not favor whites. One study of panel (three or more interviewers) interviews found that the effects of rater race and applicant race were small but that the racial composition of the panel had important practical implications in that over 20 percent of decisions would change depending on the racial composition of the interview panel. Black raters evaluated black applicants more favorably than white applicants only when they were on a predominantly black panel.[105]

Age Discrimination

Although the research indicates that older applicants generally receive lower evaluations than do younger applicants, this effect is influenced by the type of job in question, interviewer characteristics, and the content of the interview questions (i.e., traits versus

qualifications). The evidence for age bias is mixed and suggests that, as in gender bias, age bias might be largely determined by the type of job under study.[106]

Disability Discrimination

Few studies have examined bias against disabled applicants. The evidence that exists suggests that some disabled applicants receive lower hiring evaluations but higher attribute ratings for personal factors such as motivation. Before any conclusions about disability bias can be made, more research needs to be conducted that examines the nature of the disability and the impact of situational factors, such as the nature of the job. (See Chapter 3 for a discussion of the ADA.)[107]

How Can We Improve the Validity of Interviews?

Some interviewers, no doubt, are guilty of one or more of the discriminatory biases described earlier. Employers should examine their interview process for discriminatory bias, train interviewers about ways to prevent biased inquiries, provide interviewers with thorough and specific job specifications, structure the interview around a thorough and up-to-date job analysis, and monitor the activities and assessments of individual interviewers.

Use expat managers to develop and conduct expat interviews

Many multinational corporations use successful overseas managers to develop and conduct interviews for the selection of managers for international assignments. These managers tend to understand the major requirements of such jobs better than managers who have no overseas experience. Many U.S. companies, including Ford, Nestlé, Procter & Gamble, Texaco, and Philip Morris, credit improvements in their expatriate placements to their interviewing processes, which involve experienced and successful expatriates who have had experience in the same jobs to be filled.

The **physical environment** for the interviews should be maintained consistently by providing a standardized setting for the interviews. The conditions surrounding the interview might influence the decision-making process; therefore, extraneous factors such as noise, temperature, and interruptions should be controlled. Some companies use computer interviewing to standardize the interview process and reduce costs.

There is a great need for interviewer training. The previous discussion about the decision-making process indicates that interviewers need to be trained regarding how to evaluate job candidates, what criteria to use in the evaluation, how to use evaluation instruments, and how to avoid common biases and potentially illegal questions.

Johnson's Wax found that most interviewers had made their decisions about applicants after only 5 minutes. It trained its people to withhold judgment and gather information free of **first-impression bias.** Companies should use workshops and group discussions to train interviewers how to do the following:

1. *Use job information*: understand job requirements and relate these requirements to the questioning content and strategy.
2. *Reduce rating bias*: practice interviewing and provide feedback and group discussion about rating errors.
3. *Communicate effectively*: develop a rapport with applicants, "actively listen," and recognize differences in semantics.

The training should focus on the following:

1. Use of interview guides and outlines that structure the interview content and quantitatively rate applicant responses.
2. Exchange of information that focuses on relevant applicant information and provides applicants with adequate and timely information about the job and company.

Interview content

The content of the interview determines what specific factors are to be evaluated by the interviewers. The following are general suggestions based on legal and practical concerns; more specific content guidelines should be based on the specific organization and the relevant state and local laws.

1. Exclude traits that can be measured by more valid employment tests: for example, intelligence, job aptitude or ability, job skills, or knowledge.
2. Assess personality, motivational, and interpersonal factors that are required for effective job performance. These areas seem to have the most potential for incremental validity after GMA or knowledge-based tests. Use interview assessment in conjunction with standardized inventories such as a FFM instrument or the 16PF to assess relevant traits (e.g., Extraversion, Emotional Stability, and Conscientiousness for managerial jobs). Interviewers should assess only those factors that are specifically exhibited in the behavior of the applicant during the interview and that are critical for performance on the job to be filled. Don't place too much weight on interviewee anxiety.
3. Match interview questions (content areas) with the job analysis data for the job to be filled and the strategic goals of the organization.
4. Avoid biased language or jokes that may detract from the formality of the interview, and avoid inquiries that are not relevant to the job in question.
5. Limit the amount of preinterview information to information about the applicants' qualifications and clear up any ambiguous data. While knowledge of test results, letters of reference, and other sources of information can bias an interview, it is a good strategy to seek additional information relevant to applicants' levels of KASOCs.
6. Encourage note taking; it enhances recall accuracy.
7. Be aware of candidate impression management behaviors.

The format suggestions deal with how the interview content is structured and evaluated. These suggestions describe different types of interview procedures and rating forms for standardizing and documenting interviewer evaluations.

Interview questions are intended to elicit evaluation information; therefore, rating forms are recommended in order to provide a systematic scoring system for interpreting and evaluating information obtained from applicants. Based on the job analysis, the specified content of the interview, and the degree of structure for the procedure, rating forms should be constructed with the following features. First, the ratings should be behaviorally specific and based on possible applicant responses exhibited during the interview. Second, the ratings should reflect the relevant dimensions of job success and provide a focused evaluation of only the factors required for job performance. Third, the ratings should be based on quantitative rating scales that provide a continuum of possible responses. These anchors provide examples of good, average, and poor applicant responses for each interview question. The use of anchored rating forms reduces rater error and increases rater accuracy. This approach, using specific, multiple ratings for each content area of the interview, is preferred to using an overall, subjective suitability rating that is not explicitly relevant to the job. Figure 6-10 presents an example of an actual rating form.

What Are Major Types of Interviews?

A variety of interview formats are used today, but most interviews are not standardized. While this lack of standardization has contributed to low reliability and validity of both overall interview decisions and the decisions of individual interviewers, improvements in the effectiveness of the procedure have been made based on the following types of interview formats.

Interview formats

Structured interviews range from highly structured procedures to semistructured inquiries. A highly structured interview is a procedure whereby interviewers ask the same questions of all candidates in the same order. The questions are based on a job analysis and are reviewed for relevance, accuracy, ambiguity, and bias. A semistructured interview provides general guidelines, such as an outline of either mandatory or suggested questions, and recording forms for note taking and summary ratings. In contrast, the traditional, unstructured interview is characterized by open-ended questions that are not necessarily based on or related to the job to be filled. Interviewers who use either of the structured

Figure 6-10 Sample Situational Interview Questions

1. A customer comes into the store to pick up a watch he had left for repair. The repair was supposed to have been completed a week ago, but the watch is not back yet from the repair shop. The customer is very angry. How would you handle the situation?
   - 1 (low) Tell the customer the watch is not back yet and ask him to check back with you later.
   - 3 (average) Apologize, and tell the customer that you will check into the problem and call him or her back later.
   - 5 (high) Put the customer at ease and call the repair shop while the customer waits.[a]
2. For the past week you have been consistently getting the jobs that are the most time consuming (e.g., poor handwriting, complex statistical work). You know it's nobody's fault because you have been taking the jobs in priority order. You have just picked your fourth job of the day and it's another "loser." What would you do?
   - 1 (low) Thumb through the pile and take another job.
   - 2 (average) Complain to the coordinator, but do the job.
   - 3 (high) Take the job without complaining and do it.[b]

[a]Source: Jeff A. Weekley and Joseph A. Gier, "Reliability and Validity of the Situational Interview for a Sales Position," *Journal of Applied Psychology* 3 (1987), pp. 484–487. American Psychological Association. Reprinted with permission. See also:
[b]Source: Gary P. Latham and Lise M. Saari, "Do People Do What They Say? Further Studies on the Situational Interview," *Journal of Applied Psychology* 4 (1984), pp. 569–573.

interview procedures standardize the content and process of the interview, thus improving the reliability and validity of the subsequent judgments. Structured interviews are typically behavioral or situational (or both).

Group/panel interviews consist of multiple interviewers who independently record and rate applicant responses during the interview session. With panel interviews, multiple ratings are combined usually by averaging across raters. The panel typically includes the job supervisor and a personnel representative or other job expert who helped develop the interview questions. As part of the interview process, the panel reviews job specifications, interview guides, and ways to avoid rating errors prior to each interview session. Procter & Gamble uses a minimum of four interviews to be filled. The CIA uses a minimum of three interviews for each job candidate. The use of a panel interview reduces the impact of idiosyncratic biases that single interviewers might introduce, and the approach appears to increase interview reliability and validity. Many team-based production operations use team interviews to add new members and select team leaders. In general, there is greater validity in interviews that involve more than one interviewer for each job applicant. Two approaches to interviewing with excellent track records when they make up a structured interview are situational and behavioral interviews.

Situational interviews require applicants to describe how they would behave in specific situations. The interview questions are based on the critical incident method of job analysis, which calls for examples of unusually effective or ineffective job behaviors for a particular job (see Chapter 4). For situational interviews, incidents are converted into interview questions that require job applicants to describe how they would handle a given situation. Each question is accompanied with a rating scale, and interviewers evaluate applicants according to the effectiveness or ineffectiveness of their responses.

The Palm Beach County, Florida, school board asked the following question of all applicants for the job of high school principal: "Members of the PTA have complained about what they regard as overly harsh punishment imposed by one teacher regarding cheating on an exam. How would you handle the entire matter?" Another question had to do with a teacher who was not complying with regulations for administering standardized tests. The candidate was asked to provide a sequence of actions to be taken regarding the situation. The situational approach may be highly structured and may include an interview panel. In the case of Palm Beach County, three principals trained in situational interviewing listened to applicants' responses, asked questions, and then made independent evaluations of each response. The underlying assumption is that applicants' responses to the hypothetical job situations are predictive of what they would actually do on the job. This technique improves interviewer reliability and validity.

Behavioral interviews ask candidates to describe actual experiences they have had in dealing with specific, job-related issues or challenges. Behavioral interviewing may involve probing beyond the initial answer. At GM's Saturn plant, employees are first asked

to describe a project in which they participated as group or team members. Probing may involve work assignments, examples of good and bad teamwork, difficulties in completing the project, and other related projects.

For example, to test analytical skills, some possible behavioral questions are

1. Give me a specific example of a time when you used good judgment and logic in solving a problem.
2. Give me an example of a time when you used your fact-finding skills to solve a problem.
3. Describe a time when you anticipated potential problems and developed preventive measures.
4. What steps do you usually follow to study a problem before making a decision?

The "Bottom Line" on Interview Validity

Behavioral interviews have higher validity than situational interviews

While situational interviews are valid, the behavioral interviewing approach where candidates describe actual experiences or accomplishments with important job-related situations has been shown to be reliably more valid, particularly when reported achievements or accomplishments are verified or validated.[108] So, a **"high-validity" interview should be structured with behavioral questions derived from a job analysis and involving more than one trained interviewer using a structured interview rating form.** Interviewers should keep the interview as standardized as possible and derive quantitative ratings on a small number of job-related (and observable) dimensions (e.g., ability to communicate, interpersonal skills). Ratings of abilities such as GMA that can be assessed with a standardized test should be avoided (just use the test).

More companies are taking advantage of a cost-effective way to get multiple assessments of job candidates under standardized conditions by videotaping the interviews and then circulating them to evaluators, who can be anywhere when reviewing the interviewing and doing the evaluations.[109] If this cannot be done, the use of three and preferably more independent (and qualified) interviewers will probably get you comparable validity to the "high validity" approach just described.

The "high-validity" interview

Interview data should not be overemphasized but appropriately weighed with other valid information. When done as recommended, interviews can contribute to the prediction of job performance over and above tests of GMA personality tests and other measures of personal characteristics and accomplishments.

COMBINING DATA FROM VARIOUS SELECTION METHODS

A number of valid selection procedures have been described in this chapter. BA&C, the consulting firm working with Wackenhut Security, recommended an accomplishment record for its supervisory jobs, which could be completed online, followed by reference checks and a background check. Applicants also could complete an online "in-basket" performance test. The next step involved web-camera interviews between assessors and candidates, followed by a detailed behavioral interview.

But how should the data from the different selection methods be combined so that a final decision can be made regarding the applicants to be selected? As discussed earlier, most decisions are based on a "clinical" or "holistic" analysis about each candidate after reviewing assessments from several different sources. With the clinical approach, there is no formal method of weighing scores on the various selection methods or sources. Another method is to weigh scores from each approach equally after standardizing the data (standardizing each score as a deviation scores above or below the mean on any given method). Each applicant would receive a standard score on each predictor, the standard scores would be summed, and candidates would then be ranked according to the summed scores. A better approach calls for rank ordering candidates on each method and then averaging the

ranks for each candidate (the top candidate would have the lowest average rank). Another superior approach, which can be combined with the standardizing and rank ordering, is to weigh scores based on their empirical validity; that is, the extent to which each method is correlated with the criterion of interest (e.g., sales, performance, turnover). An alternative approach to the use of reported validities is to rely on expert judgment regarding the weight that should be given to each selection method. Experts could review the content and procedures of each of the methods and give each a relative predictive weight that is then applied to applicant scores.

Weigh data based on validity of method

One of the "discrepancies" between research and practice is the clear academic finding that **"actuarial"** or "statistical" decision making is superior to "clinical" or "holistic" prediction. Actuarial prediction means you should derive a formula that weighs information based on the relative validity of the different sources and how each candidate performed on that source (after standardizing the data). Next, a score is derived for each candidate based on this formula. This "actuarial" approach is superior to studying a lot of information (some valid, some not so much) and then making an overall "clinical" assessment (or prediction). If you can't use validity coefficients, using an average rank ordering process (across methods) is recommended and is superior to "clinical" judgment.[110]

Use "actuarial" prediction not "holistic"

BA&C conducted a large scale, criterion-related validity study and derived weights based on the validity of each of the data sources. Structured, behavioral interviewing for only the top candidates was recommended based on the number of positions they had to fill. This multiple-step process saved time and money. Most companies that use a variety of different instruments follow a similar procedure by initially using the least expensive procedure (e.g., GMA and non-cognitive measures, biodata) and then using a set of procedures, such as performance tests, for those who do well in the first round. These companies perform interviews only on the top scorers from the second phase of testing. The CIA, the FBI, numerous insurance companies, and a number of the most prestigious graduate business schools follow a similar procedure. The Wharton School at the University of Pennsylvania does initial screening on the basis of the GMAT and undergraduate performance. The school then requests answers to lengthy essay test questions. If the student survives this hurdle, several faculty members conduct interviews with the student.

Interviewing, especially in this context, is perhaps the most important of the selection options for assessing the person–organization fit. Google, for example, interviews job applicants several times by as many as 20 interviewers. Toyota (USA) conducts a formal interview for its Georgetown, Kentucky, factory jobs. The interview results are combined with assessment center data, a work sample, and an aptitude test. **The most effective selection systems integrate the data from the interview with other sources and weigh the information using the person–organizational fit model.** Take note also that self-report personality measures are more prone to faking than structured interviews designed to measure the same (and job-related) personality traits factors.

Connecticut v. Teal

What are the legal implications of this multiple-step process? In the *Connecticut v. Teal* case (see Chapter 3), Ms. Teal was eliminated from further consideration at the first step of a multiple-step selection process and claimed she was a victim of Title VII discrimination. The Supreme Court said that even if the company actually hired a disproportionately greater number of minorities after the entire selection process, the **job relatedness** *of that first step* must be determined because this was where Ms. Teal was eliminated.

One excellent example of the effectiveness of using multiple measures to predict is a study that focused on predicting college student performance.[111] Scores from a biographical instrument and a situational judgment inventory (SJI) provided incremental validity when considered in combination with standardized college-entrance tests (i.e., SAT/ACT) and a measure of Big-Five personality factors. Also, racial subgroup mean differences were much smaller on the biodata and SJI measures than on the standardized tests and college grade point average. Female students outperformed male students on most predictors and outcomes with the exception of the SAT/ACT. The biodata and SJI measures clearly showed promise for selecting students with reduced adverse impact against minorities.

What Is Individual Assessment?

Individual assessment (IA) is a very popular approach for selecting managers although there has been little research to determine validity. This approach is almost always based on an overall assessment provided by one or more psychologists. The IA is based on information from several sources discussed in this chapter. A lengthy interview and psychological testing, often using projective measures, are almost always involved. The Tribune Company, for example, often used the services of a company that (for $3,500 per candidate) provides a psychological report on the candidate's prospects based on scores on the 16PF personality test (which measures the Big-Five factors and sub-factors), a GMA test, and a detailed interview with a psychologist who is bases his or her assessment on some prototype of the "ideal" manager. While the psychologist for this company could have used some statistical model for the final assessment based on the relative validity of the various sources of information about the candidates, like almost all IA, the report is based on a "holistic" or clinical assessment of the candidate as a "whole" where the psychologist studies all the information and then writes the report based on his or her own impression.

"Holistic" approach not recommended

This is another example of the discrepancy between research and practice. The research shows to use a statistical or actuarial model based on the relative validity of the various sources of information. An excellent review of this approach to assessment was very critical of the method and concluded that "the holistic approach to judgment and prediction has not held up to scientific scrutiny."[112]

Setting cut-off scores

Another issue is where you set the cutoff score in a multiple-cutoff system such as that recommended by BA&C. Where, for example, do you set the cutoff score for the paper-and-pencil tests in order to identify those eligible for further testing? Unfortunately, there is no clear answer to this important question. If data are available, cutoff scores for any step in the process generally should be set to ensure that a *minimum* predicted standard of job performance is met. If data are not available, cutoff scores should be set based on a consideration of the cost of subsequent selection procedures per candidate, the legal defensibility of each step in the process (i.e., job relatedness), and the adverse impact of possible scores at each step. As discussed in Chapters 3 and 4, cutoff scores can be at the center of litigation if a particular cutoff score causes adverse impact. As discussed earlier, the City of Chicago lost a Title VII lawsuit because the particular cutoff score used for the firefighters exam caused adverse impact and was not shown to be "job related."[113] Recall the discussion in Chapter 3 about the plaintiff's opportunity to present evidence and testimony for an alternative method with comparable validity and less adverse impact. The lower cutoff score has been offered successfully as the alternative method. Where the hiring of people who turn out to be ineffective is unacceptable, as, for example, in armed security positions at airports, the setting of a higher (more rigorous) cutoff score is clearly necessary.

PERSONNEL SELECTION FOR OVERSEAS ASSIGNMENTS*

One expert on expatriate assignments tells the story of a major U.S. food manufacturer who selected the new head of the marketing division in Japan. The assumption made in the selection process was that the management skills required for successful performance in the United States were identical to the requirements for an overseas assignment. The new director was selected primarily because of his superior marketing skills. Within 18 months, his company lost 89 percent of its existing market share.[114]

What went wrong? The problem may have been the criteria that were used in the selection process. The selection criteria used to hire a manager for an overseas position must focus on more facets of a manager than the selection of someone for a domestic position. The weight given to the various criteria also may be different for overseas assignments. Besides succeeding in a job, an effective expatriate must adjust to a variety of factors: differing job responsibilities even though the same job title is used, language and cultural barriers that make the training of local personnel difficult, family matters such as spouse employment

*Stephanie Thomason assisted in the preparation of this section.

and family readjustment, simple routine activities that are frustrating in the new culture, and the lack of traditional support systems such as religious institutions or social clubs. The marketing head in Japan, for example, spent considerable time during the first 6 months of his assignment simply trying to deal with family problems and to adjust to the new environment. This experience is hardly unique. As discussed in Chapter 2, expatriate selection is a real challenge, often cited by senior human resource managers as one of the most likely causes of expatriate assignment failure.[115] One survey of 80 U.S. multinational corporations found that over 50 percent of the companies had expatriate failure rates of 20 percent or more.[116] The reasons cited for the high failure rate were as follows (presented in order of importance): (1) inability of the manager's spouse to adjust to the new environment, (2) the manager's inability to adapt to a new culture and environment, (3) the manager's personality or emotional immaturity, (4) the manager's inability to cope with new overseas responsibilities, (5) the manager's lack of technical competence, and (6) the manager's lack of motivation to work overseas. Obviously, some of these problems have to do with training and career issues. Figure 6-11 presents an often-cited model of expatriate selection, which identifies job and personal categories of attributes of expatriate success.

Expatriate failures related to selection

Several of the factors listed previously concern the process of selecting personnel for such assignments. The food manufacturer placed almost all the decision weight on the technical competence of the individual, apparently figuring that he and his family could adjust or adapt to almost anything. In fact, we now know that adjustment can be predicted to some extent, and that selection systems should place emphasis on adaptability along with the ability to interact well with a diverse group of clients, customers, and business associates. Surprisingly, few organizations place emphasis on so-called relational abilities in the selection of expatriates. One review found that despite the existence of useful tests and questionnaires, "many global organizations do not use them extensively because they can be viewed as overly intrusive."[117] Studies involving the **Big Five** or FFM show better cross-cultural adjustment with higher scores in "Openness to Experience" and stronger performance with high "Conscientiousness" scores.[118] One meta-analysis of 30 studies and over 4,000 respondents found that in addition to conscientiousness, extraversion, emotional stability, and agreeableness predict expatriate job performance. While openness to experience did not predict job performance, additional factors such as cultural sensitivity and local language ability did.[119]

Relational ability

The FFM and expatriate success

One study of expatriates working in Japan, Hong Kong, and Korea found that high levels of emotional stability and openness to experience had more to do with who would succeed or fail than technical knowledge. Doing a better job identifying expatriates' successes was very important for the firm under study. The researchers estimated that the cost of failure was over $150,000 per expatriate.[120]

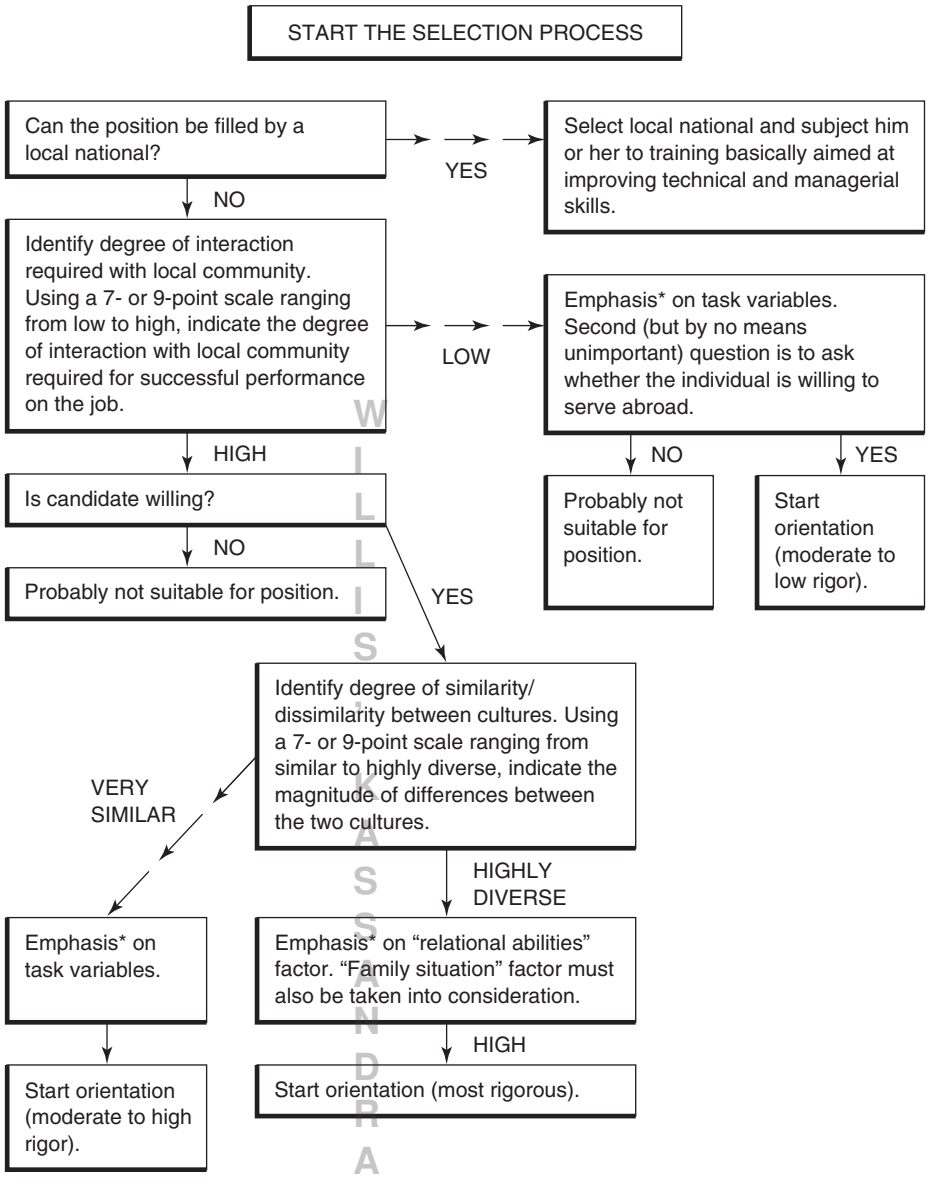
Of course, one critical question that must first be addressed is whether a corporation would be better off hiring someone from within the host country. Figure 6-12 presents a decision model that addresses this option. If the answer to this question is no, the model provides a chronology of the questions to be answered in the selection of an expatriate. If

Figure 6-11 **Categories of Attributes of Expatriate Success**

Job Factors	Relational Dimensions	Motivational State	Family Situation	Language Skills
Technical skills	Tolerance for ambiguity	Belief in the mission	Willingness of spouse to live abroad	Host country language
Familiarity with host country and HQ operations	Behavioral flexibility	Congruence with career path	Adaptive and supportive spouse	Nonverbal communication
Managerial skills	Nonjudgmentalism	Interest in overseas experience	Stable marriage	
Administrative competence	Cultural empathy and low ethnocentrism	Interest in specific host country culture		
	Interpersonal skills	Willingness to acquire new patterns of behavior and attitudes		

Source: S. Ronen, *Training the International Assignee: Training and Career Development*, 1st ed. (San Francisco: Goldstein, 1989). See also J. Chew, "Managing MNC Expatriates through Crises: A Challenge for International Human Resource Management," *Research and Practice in Human Resource Management,* 12 (2) (2004), pp. 1–30.

Figure 6-12
Model of the Selection Process for Overseas Assignments



Source: Reprinted from R. L. Tung, "Selection and Training for Overseas Assignments," *Columbia Journal of World Business* 16 (1981), pp. 68–78. Reprinted with permission from Elsevier.

the answer is yes, the decision makers must be aware of any applicable host laws regarding personnel selection. In Poland and Sweden, for example, prospective employees must have prior knowledge of any testing and can prohibit the release of testing data to the company. Many European countries require union participation in all selection decisions for host nationals. Thus, companies may find that hiring host nationals is more problematic than going the expatriate route. Assuming that the host option is rejected, what steps should be followed to make better selection decisions about expatriates? Let us examine some organizations that select large numbers of expatriates successfully.

The Peace Corps has only about a 12 percent turnover rate (i.e., people who prematurely end their assignments). Of the 12 percent, only 3 to 4 percent are attributed to selection errors. The Peace Corps receives an average of 5,000 applications per month. The selection process begins with an elaborate application and biographical data form that provides information on background, education, vocational preferences, and volunteer activity in the past. Second, the applicant must take a placement test to assess GMA and language aptitude. Third, college or

high school transcripts are used for placement rather than screening. The fourth step requires up to 15 references from a variety of sources. Although the general tendency among references is to provide positive views of candidates, one study found that for sensitive positions such as the Peace Corps volunteer, references often provide candid comments about applicants. The final step is an interview with several Peace Corp representatives. During the interview process, the candidate is asked about preferred site locations and specific skills as well as how he or she would deal with hypothetical overseas problems. An ideal candidate must be flexible and tolerant of others and must indicate a capacity to get work done under adverse conditions. The interviews also provide Peace Corps staff with details concerning the candidate's background and preferences so that appropriate work assignments may be determined.

Based on the preceding four sources of information, the screeners assess a candidate using the following questions: (1) Does the applicant have a skill that is needed overseas, or a background that indicates he or she may be able to develop such a skill within a 3-month training period? This question is designed to match the candidate with a job required by a foreign government, such as botanist, small business consultant, or medical worker. (2) Is the applicant personally suited for the assignment? This question focuses on personality traits such as adaptability, conscientiousness, and emotional stability.

Weights for Expatriate Selection

Structure reproducer selection

The weight to be given to expatriate selection factors differs as a function of the position to be filled. For example, a position that has an operational element requiring an individual to perform in a preexisting structure does not require strong interpersonal skills. However, a "structure reproducer," an individual who builds a unit or department, does need strong interpersonal skills. Thus, the selection system should focus on the cultural environment, job elements, and individual talents. The weights given to the various criteria should be determined by the individual job. A job analysis would be helpful in this regard. This system is exemplified by Texas Instruments (TI), a manufacturer of electronics and high-technology equipment based in Dallas. In seeking expatriates for start-up ventures, the company focuses on such issues as an individual's familiarity with the region and culture (environment), specific job knowledge for the venture (job elements), knowledge of the language spoken in the region, and interpersonal skills. TI uses several methods to make assessments on these dimensions, including the Five-Factor Model.

Many companies emphasize the "manager as ambassador" approach since the expatriate may act as the sole representative of the home office. IBM and GE, for example, select people who best symbolize the esprit de corps of the company and who recognize the importance of overseas assignments for the company.

Use a realistic job preview for expatriate assignments

A review of the most successful systems for selecting expatriates provides a set of recommendations for a selection system. First, potential expatriates are identified through posted announcements, peer and/or superior nominations, or performance appraisal data. Second, promising candidates are contacted and presented with an overview of the work assignment. A **realistic job preview** would be ideal at this stage. Third, applicants are examined using a number of selection methods, including paper-and-pencil and performance tests. A growing number of companies now use standardized instruments to assess personality traits. The 16PF, for example, has been used for years to select overseas personnel for the U.S. Department of State and is used by some U.S. companies and executive search companies that specialize in expatriate assignments. Although **relational ability** is considered to be a major predictor of expatriate success, the one available survey on the subject found that only 5 percent of companies were assessing this ability through a formal process (e.g., paper-and-pencil tests, performance appraisals).

Successful expats are ideal interviewers

After a small pool of qualified candidates is identified, candidates are interviewed and the best matches are selected for the assignment. Successful expatriates are ideal as interviewers. Our coverage of employment interviews provides recommendations for enhancing the validity of these interview decisions. Do the more rigorous selection systems result in a higher rate of expatriate success? The answer is clearly "yes."

Two tests that have been shown to be useful (and valid) are the **Global Assignment Preparedness Survey,** which assesses candidates on six dimensions, including cultural flexibility, and the **Cross-Cultural Adaptability Inventory,** which focuses on the ability to adapt to new situations and interact with people different from oneself.[121]

SELECTION IN OTHER COUNTRIES

Background checks on job candidates are far less common in Europe and Asia and have more restrictions. For most employment situations, job applicants must grant consent for a background check, and most would actually refuse to grant this consent. This should not be such a concern to U.S. companies doing business in Europe. Negligent hiring is not accepted as a legal theory in Europe and the value of conducting detailed background checks is rather dubious.[122]

The use of employment tests in other countries of the world varies considerably as do the government regulations regarding the use of tests. Turning first to Asian countries, Korean employers report the use of employment tests extensively and more than any other country.[123] These tests tend to be written examinations covering English language skills, common sense, and knowledge of specific disciplines. A smaller percentage of Japanese companies use employment tests. Some Japanese companies use the **Foreign Assignment Selection Test** (FAST) to identify Japanese who are more likely to be successful expatriates in the United States. The FAST assesses cultural flexibility, sociability, conflict resolution style, and leadership style. Within Japan, however, most people are hired directly from the universities, and the prestige of the university attended is a major criterion for selection purposes. A survey of companies in Hong Kong and Singapore revealed little use of employment tests, but there are a growing number of U.S. companies that have opened offices in Hong Kong. Aside from some use of clerical and office tests (e.g., typing), only two companies from these countries indicated use of any personality, cognitive ability, or related tests. Finally, recent evidence indicates that China makes extensive use of employment testing, contrary to previous research.[124]

European countries have more controls on the use of tests and other methods for selection, but there is considerable variability in usage. Due to the power of unions in most European countries, employers have more restrictions on the use of tests for employment decisions, compared to the United States. A wide variety of employment tests appear to be used in Switzerland, including graphology, but in Italy selection tests are heavily regulated. In Holland, Sweden, and Poland, job applicants have access to all psychological test results and can choose to not allow the results to be divulged to an employer.[125]

Several surveys have given us clues about selection methods in England. One survey found that more than 80 percent of companies in England do some type of reference check and another found that almost 40 percent had used personality tests and 25 percent had used cognitive ability tests to assess manager candidates.[126] About 8 percent of the surveyed firms in England reported using cognitive ability tests to select managers.

In general, there is wide variation in the use of employment tests outside the United States. While some countries have restricted the use of tests (e.g., Italy), their use appears to be far more extensive in others (e.g., China, Korea). The United States and England appear to be major centers for research and development of employment tests. Japanese companies make extensive use of testing for their U.S. plants as well as for their expatriates.[127] Their Nissan plant in Tennessee relies on team assessment using a structured interview and a battery of cognitive ability tests to select new team members.

U.S. HRM specialists considering the use of tests outside of the United States to hire employees must be very familiar with laws and regulations within the country where the testing is being considered. These laws, regulations, and collective bargaining issues are very different across countries.

THE BOTTOM LINE ON STAFFING

Figure 6-13 presents a chronology of steps that should be followed based on solid research and legal considerations. You should note that effective selection requires effective recruiting. That recruiting should be done only when the organization has determined which KASOCs or competencies are required to execute strategic goals.[128]

Figure 6-13 The Bottom-Line Chronology on Staffing

1. DEFINE THE JOB WITH A FOCUS ON JOB SPECIFICATIONS (COMPETENCIES) COMPATIBLE WITH STRATEGIC GOALS AND EXECUTING THOSE GOALS

Action: Re-do job descriptions/specifications or competencies.
Define critical KASOCs/competencies.

2. RECRUIT FROM A BROAD POOL OF CANDIDATES

Action: Lower selection ratio (increase number of qualified applicants for key positions) through better and more focused recruiting; for managerial positions, emphasize internal talent.
Increase pool of qualified minorities.

3. USE VALID INITIAL SCREENING DEVICES

Action: Develop or purchase most valid and most practical screening devices with the least adverse impact.
Refer to Mental Measurements Yearbook (www.unl.edu/Buros) for test reviews.
If using Validity Generalization (VG) research to validate, make certain the VG study has sufficient detail to show similar jobs were studied.
Where more than one valid selection procedure is available, equally valid for a given purpose, use the procedure which has been demonstrated to have the lesser adverse impact.
Use more than one method to assess job-related traits/competencies (e.g., self-reported inventories and interviews).
Develop weighting scheme (an actuarial predictive model) for competencies and the information sources that purport to measure them (including interview data).

4. DO BACKGROUND/REFERENCE CHECKS

Action: Develop performance-based reference checking focused on KASOCs/competencies.

5. USE BEHAVIORAL INTERVIEWING TECHNIQUE WITH STRUCTURED FORMAT OR INDEPENDENT MULTIPLE INTERVIEWERS ASKING BEHAVIORAL QUESTIONS

Action: Develop questions to assess KASOCs/competencies.
Train interviewers on valid interviewing and legal issues.
Derive a scoring system for interviews regardless of format.

6. USE WEIGHTING SCHEME FOR INFORMATION

Action: Derive weighting scheme based on relative importance of KASOCs/competencies and/or relative validity of the sources of information on each critical KASOC/competency.
Use "actuarial" not clinical or holistic method for ranking candidates.

7. EXTEND AN OFFER

Action: Offer should be in writing with the facts of the offer; train employees to avoid statements regarding future promotions, promises of long-term employment, etc.

Adapted from: W. F. Cascio and H. Aguinis, "Test Development and Use: New Twists on Old Questions," *Human Resource Management,* 44 (2005), pp. 219–236.

SUMMARY

Personnel selection continues to be a critical HRM responsibility. A number of commonly used tests and other assessment methods have been reviewed. While GMA or cognitive ability tests are among the most valid measures, they also frequently result in adverse impact against minority groups. Conversely, many personality tests are safe from legal problems because they typically have no adverse impact, yet are less valid. These noncognitive measures are clearly less valid than GMA in the prediction of overall job performance. It is clear that job-related personality/motivational constructs should be assessed but that multiple approaches to their measurement should be used (e.g., an inventory and an interview) for greater reliability and validity in the measurement of these constructs. The use of compound traits (more job-related, targeted noncognitive measures) will probably increase validity.

Many companies also use preemployment drug tests. These tests are generally legal to use, but there are differences from state to state. There is evidence that drug tests will screen out less-effective employees. Reference checks may not be a particularly valid selection device; still, court decisions regarding negligent hiring lawsuits indicate that employers should do their best to check applicant references. Many companies now use

integrity tests because of the restrictions on polygraph testing. Despite some political activity to amend the polygraph law by including a federal ban on these tests as well, research on these tests seems to support their use.

Assessment centers are ideal for managerial selection

Assessment center and performance testing results are valid, job related, more legally defensible, but certainly more expensive than other selection techniques, including the employment interview. Assessment centers are ideal for managerial jobs with both internal and external candidates. There is evidence that structured, behavioral interviewing conducted by more than one interviewer (the "high-validity" interview) can increase the validity of interviews unless unstructured interviews are conducted independently by three or more interviewers. Most companies use a variety of selection procedures, proceeding through the process in the order described in the model in Figure 6-1. But few organizations combine the information using an actuarial or statistical model or expert weighting model, which enhances the accuracy of decision making. Unfortunately, most companies gather information from several sources (e.g., application blanks, cognitive, and personality tests) and apply a subjective and unreliable weighting system to determine the rank of candidates for the positions to be filled. Almost all companies use an employment interview at some point in the selection process. These companies also tend to place entirely too much weight on the results of an unstructured interview that does not approach the characteristics of a "high-validity" interview.

Use behavioral interviewing

Use actuarial model

The accuracy of interview decisions is limited by the information-processing capabilities of interviewers. Factors such as the characteristics of the applicant, the interviewer, and the situation can influence and distort the decision-making process, resulting in less-than-optimal interview decisions. Because employment interviews entail complex decision-making activities, interviewers often try to simplify that process and, in doing so, bias their decisions. This inherent bias poses both legal and practical implications for management. Overall organization performance can be affected because interviewer bias reduces the probability of selecting the highest-performing candidates.

The administrative guidelines described in this chapter help ensure that the validity of the interview is maximized while interviewer bias is minimized. In turn, the procedural guidelines define both the content and the method of the interview inquiry, providing a means of improving the overall effectiveness of the interview procedure. A final dilemma facing organizations that use the interview as a selection tool continues to be the issue of **"functional utility":** What is the unique contribution of the interview in the employment decision? This is a practical assessment of the usefulness of the interview based on a determination of which information is best collected through the interview process and whether interviewer decisions based on that information are consistent and accurate. In order to achieve any functional utility from the interview, organizations must evaluate their overall selection procedures and determine (1) what factors are best and most consistently evaluated during the interview and (2) whether other selection procedures can measure those identified factors as well as or better than the interview. Organizations also should focus on the purpose of selection interview. Interviews that attempt to assess candidate "fit" while simultaneously recruiting the candidate usually fail at both.

Try to match the person with the job

The most effective personnel selection systems place a great emphasis on the interaction of the person and the organization in the prediction of effectiveness. The "matching" model presented in Chapter 5, for example, calls for an assessment of the applicant in the context of both job and organizational characteristics and a realistic assessment of the organization and the job by the applicant. This "matching" model is particularly effective in "high-involvement organizations" where employees have more latitude in the workplace. As stated at the outset of this chapter, the tools used for selection should ideally be the most valid for the particular KASOCs or competencies most important for strategic execution. This is the optimal "matching" model.

The "fairness factor"

Labor attorney Rita Risser recommends that the "fairness factor" be kept in mind by line managers making hiring decisions.[129] The "fairness factor" is expressed in five questions that should be asked in every hiring decision: (1) Am I basing decisions solely on job-related criteria? (2) Am I treating people consistently? (3) Am I following organizational policy? (4) Am I communicating accurately and honestly? and (5) Should I consult with an HR specialist or a legal expert? Ms. Risser maintains that managers who follow the "fairness factor" are more likely to make selection decisions that are free from bias or

the perception of bias. Of course, the answers to the first four fairness questions should be "yes" and the importance of the answer to the fifth question about consulting an expert really depends on how knowledgeable the decision maker is about the legal implications of the action. At the most basic level, managers should know that it is either unlawful or potentially unlawful to do any of the following:

1. Base decisions on characteristics such as disability, medical records, pregnancy, parental status, religion, race, sex, age, or national origin. Some states and municipalities also offer protection for sexual orientation, marital status, and other characteristics.
2. Show prejudice in recruiting or advertising for or against persons with particular protected class characteristics.
3. Request information regarding mental and physical disabilities during the interview.
4. Use methods that cannot be shown to be job related or a business necessity and that cause adverse impact.
5. Make inquiries that reveal protected class characteristics. Questions dealing with place of birth, religious affiliations, citizenship of parents, attitudes toward or histories regarding labor unions, and political views are examples of potentially troubling inquiries.

Discussion Questions

1. Are GMA or cognitive ability tests more trouble than they are worth? Given that some minorities may score lower on such tests, would it not be advisable to find some other method for predicting job success?
2. Why do you need tests of clerical ability? Couldn't you just rely on a typing test and recommendations from previous employers?
3. Under what circumstances would GMA or cognitive ability tests be appropriate for promotion decisions? Are there other methods that might be more valid?
4. If you were given a personality test as part of an employment application process, would you answer the questions honestly or would you attempt to answer the questions based on your image of the "correct" way to answer? What implications does your response have for the validity of personality testing? What does the evidence on faking show?
5. Discuss the advantages and disadvantages of performance testing and work samples. Under what circumstances would such tests be most appropriate?
6. Given that the validity of assessment centers and work samples are not substantially different than that reported for cognitive ability tests, why would an organization choose the far more costly approaches?
7. It has been proposed that students be assessed with work simulations similar to those used in managerial assessment centers. Assessments are then made on a student's competencies in decision making, leadership, oral communication, planning and organizing, written communication, and self-objectivity. What other methods could be used to assess student competencies in these areas?
8. What is stereotyping? Give examples of legal and illegal stereotypes.
9. Describe how an organization might improve the reliability and validity of the interview.
10. Contrast an unstructured interview with a situational or behavioral interview.
11. "The most efficient solution to the problem of interview validity is to do away with the interview and substitute paper-and-pencil measures." Do you agree or disagree? Explain.
12. Explain the difference between "actuarial" or statistical and "clinical" or "holistic" prediction.

WILLIS, KASSANDRA 2161TS

Part 3 Developing Human Resource Capability

WILLIS, KASSANDRA 2161TS

Chapter 7

Performance Management and Appraisal

OBJECTIVES

After reading this chapter, you should be able to

1. Understand the value and uses of performance management in organizations.
2. Know the prescriptions for more effective performance management and appraisal.
3. Define performance and distinguish between performance and correlates of performance.
4. Discuss the legal implications of performance appraisal.
5. Explain the various errors and biases in ratings and proven methods to reduce them.
6. Describe the ProMES system and report on its effects.
7. Describe the necessary steps for implementing an effective Performance Management and appraisal feedback system.

OVERVIEW

As one review concluded, "the appraisal of performance appraisal is not good."[1] A more recent review was also quite critical of this HR function and its track record, concluding that "done effectively, performance management communicates what's important to the organization, drives employees to achieve results, and implements the organization's strategy. Done poorly, performance management not only fails to achieve these benefits but can also undermine employee confidence and damage relationships."[2] While most organizations report the use of formal systems of performance management and appraisal, the majority of those express considerable dissatisfaction with them.[3]

UCLA Professor Samuel Culbert is probably the most often quoted advocate of getting rid of performance reviews. Says Professor Culbert, "a one-side-accountable, boss-administered review is little more than a dysfunctional pretense. It's a negative to corporate performance, an obstacle to straight-talk relationships, and a prime cause of low morale at work. Even the mere knowledge that such an event will take place damages daily communications and teamwork."[4] Indeed, there is considerable evidence that raters, ratees, and administrators are often dissatisfied with their performance management and appraisal systems (PM&A).

The good news is that there is sound research that points the way toward more effective PM&A from all three of these perspectives. We do not agree with Professor Culbert and others who want to do away with formal performance appraisals. We will present research-based recommendations that should make PM&A a more effective HR function.

All of the attention paid to performance appraisal is testimony to its potentially pivotal role in influencing organizational performance and effectiveness. Indeed, formal performance appraisal and multirater systems are components of **high-performance work practices** and have been linked to corporate financial performance.[5] Central to this linkage is the view that the most effective PM&A systems recognize that appraisal is not an end in itself; rather it is a critical component of a much broader set of human resource practices that are linked to business objectives, personal and organizational development, and corporate strategy.[6] **Performance management should be viewed as a "continuous process of identifying, measuring, and developing the performance of individuals and teams and aligning performance with the strategic goals of the organization."**[7]

Organizations are constantly searching for better ways to appraise performance. Sometimes these "better ways" don't work out well. Ford installed a new and controversial PM&A system called forced distribution as part of a major restructuring effort. The old Ford system resulted in such uniformly high ratings that few performance distinctions could be made among the workers and the data indicated that there were almost no ineffective workers. Forced distribution "forced" all managers to identify a certain number of ineffective workers. A few years later, Ford was settling a $100 million lawsuit that was a consequence of the use and results of the new forced distribution system. Microsoft dropped a similar PM&A system after a flood of complaints from supervisors and their subordinates. Pratt & Whitney, the jet engine division of United Technologies made significant changes in their performance appraisal and management systems in 3 consecutive years.[8]

The critical role of performance appraisal in EEO and other work-related litigation should also be emphasized, particularly in rebuttal to those who advocate getting rid of formal appraisal. **Performance appraisal is the most heavily litigated personnel practice today.** Since the legal grounds for challenging appraisal systems are expanding, litigation can be expected to increase. For example, the 2008 Supreme Court ruling in *Meacham v. Knolls Atomic Power* has placed a greater burden on employers to justify their performance appraisal decisions and practices.[9]

Meacham v. Knolls Atomic Power

As we discussed in Chapter 3, the growing diversity and aging of the workforce also increases the probability of legal and work-related difficulties. With greater proportions of women, members of minority groups, people of varying sexual orientation, employees with disabilities, and older workers in the labor force, unfairness and biases already present in appraisal systems, either real or perceived, may be magnified by greater diversity among those who evaluate performance and those who are evaluated.[10] Consequently, organizations will need to be increasingly conscientious about facilitating fairness and objectivity in appraisal practices and personnel decisions and eliminating as much subjectivity in the process as is possible.

The overall objective of this chapter is to provide recommendations for improving the effectiveness of performance management and appraisal in organizations. **There are major discrepancies between the way in which appraisal is *practiced* and the way in which experts say it *should* be done.** These discrepancies are emphasized throughout the chapter.

Major discrepancies between research and practice

There is hope for performance management and appraisal. Reviews of research, practice, and litigation related to appraisal have led to the recognition that there are some prescriptions that should be followed in order to improve the effectiveness of PM&A systems.[11] We believe that the effects of PM&A will be more positive if and when these prescriptions are followed that generally have *not* been heeded by most practitioners. The major prescriptions are

1. Precision in the definition and measurement of performance is a key element of effective PM&A.
2. The content and measurement of performance should derive mainly from internal and external customers in the context of organizational objectives.

3. Multiple raters (internal and external customers) should be used to derive ratings.
4. Incorporate a formal process for investigating and correcting the effects of situational constraints on performance.

Figure 7-1 presents an elaboration of these prescriptions, including specific recommendations subsumed under each of them.

As discussed in Chapter 1, research shows that PM&A, when done correctly, can (and does) affect organizational performance and the bottom line. Chapter 4 covers the role of performance measurement as a focus in work and job design and analysis. The role of PA for succession planning, recruitment, and downsizing is emphasized in Chapter 5. In Chapter 6, the role of PA for promotion systems and validating selection measures is emphasized. To be effective, PM&A must be a continuous process that serves to define, measure, and develop performance at the individual and the unit levels and closely links these performance measures to the strategic objectives of the organization.

PM&A is a continuous process

Performance management and appraisal practice have improved in recent years but still have a long way to go. Figure 7-2 presents a summary of findings concerning discrepancies between research and practice.

HOW DO WE DEFINE PERFORMANCE AND WHY DO WE MEASURE IT?

Despite the importance of PM&A, few organizations clearly define what it is they are trying to measure. In order to design a system for appraising performance, it is important to first define what is meant by the term **work performance.** As discussed in Chapter 6, although a person's performance depends on some combination of ability (or competency), motivation or effort, and of course the opportunity to perform, performance should be measured in terms of outcomes or results produced in the context of opportunities to perform. These outcomes or results should be closely aligned with organizational objectives. We define ***performance as the record of outcomes produced on specified job functions or activities during a specified period.***[12] For example, a trainer working for the World Bank was evaluated on her "organization of presentations," which was defined as "the presentation of training material in a logical and methodical order." The extent to which she was able to make such "methodical" presentations would be one measure of outcomes related to that function. Those outcomes were evaluated by the clients who received the training.

Obviously a sales representative would have some measure of actual sales as an outcome for the primary function of that job (i.e., sales). Customer service is a likely candidate as another important function that would have very different outcome measures for defining performance. College professors are typically evaluated on three general work

Figure 7-1
Prescriptions for Effective Performance Management and Appraisal

1. Strive for as much precision in defining and measuring performance dimensions as is feasible.
   - Define performance with a focus on valued outcomes tied to strategic goals. Where possible, use objective, countable results aligned with organizational goals.
   - If ratings are necessary for certain functions, define outcome effectiveness measures in terms of relative frequencies of outcomes (e.g., 0 to 100% of all opportunities).
   - Define performance dimensions by combining functions with aspects of value (e.g., quantity, quality, timeliness, effects on constituents, cost).
2. Link performance dimensions to meeting internal and external customer requirements.
   - Internal customer definitions of performance should be linked to external customer satisfaction.
3. Use a multi-rater system for PM&A.
4. Incorporate the measurement of situational constraints.
   - Focus attention and training on perceived constraints on performance.

Source: Adapted from H. J. Bernardin, C. Hagan, J. S. Kane, and P. Villanova, "Effective Performance Management: Precision in Measurement with a Focus on Customers and Situational Constraints," in *Performance Appraisal: State-of-the-Art Methods for Performance Management,* ed. J. Smither (San Francisco: Jossey-Bass, 1998).

Figure 7-2 Performance Management: Discrepancies between Research and Practice

Rating Content

Finding: Do not evaluate people on traits in performance appraisal.
Practice: 58% of surveyed employers still use traits as criteria.

Finding: Performance dimensions or criteria should be linked to job descriptions.
Practice: 60% of employers report strong linkage; 22% actually evaluate the linkage.

Finding: Setting precise, challenging goals results in higher performance.
Practice: 26% of managerial appraisal goals/objectives are precise.

Finding: Clearly distinguish among aspects of performance (e.g., quality, quantity).
Practice: 14% of employers distinguish aspects of value by job function or goal.

Finding: Link individual performance dimensions to specific strategic goals.
Practice: 9% actually do this; 55% make the claim.

Rating Process

Finding: Employee participation in goal setting increases motivation, commitment, and performance.
Practice: 18% of nonmanagement positions set goals; 58% of management positions allow participation.

Finding: Specific feedback focuses attention on goals.
Practice: 37% of employees indicate they received detailed feedback.

Finding: Establish tight link between goal attainment and rewards.
Practice: 41% of employees perceive a "close link" of goal attainment to rewards.

Finding: Train raters for common frame of reference (FOR).
Practice: 8% of employers use FOR; only 21% know what FOR training is.

Finding: Train raters on giving negative feedback.
Practice: 27% of employers provide such training.

Finding: Avoid training on rater error distributions—it can create other errors.
Practice: 41% of employers use rater error training.

Finding: Structured diary keeping increases reliability in rating.
Practice: 5% of companies require diary keeping by supervisors.

Finding: Train raters on cognitive errors like actor/observer bias.
Practice: 8% of employers know what this error is; 3% train on it.

Finding: Distinguish between ratings of person's characteristics and performance outcomes.
Practice: 46% of employers now rate on competencies and don't clearly distinguish between performance and ratee potential, KASOCs, or competencies.

Administrative Uses

Finding: 360-degree (or, multirater) appraisal data can reduce adverse impact in promotions.
Practice: 16% of companies that use 360-degree appraisal use it for decision making; 84% of companies rely on "top-down" appraisal for promotions.

Finding: Multirater appraisal has higher validity than "top down appraisal."
Practice: Less than 5 percent of companies use multirater appraisal for decision making.

Rating Results

Finding: Audit data for adverse impact against protected classes (including age).
Practice: 24% of companies do this annually; 63% have never done it.

Finding: Evaluate particular rater tendencies (e.g., ratings by ethnicity, gender, age, leniency, other rating errors).
Practice: 15% of companies calculate rating data by rater.

Finding: Reward raters for rating process adherence (e.g., precise criteria, good differentiation).
Practice: 27% of companies include performance management practices as critical component of managers' jobs.

Finding: Assess individual performance levels as related to aggregated, strategic goals.
Practice: 24% actually do this in any way; 58% make the claim.

Source: Adapted from H. J. Bernardin, "Survey of HR Practice: More Evidence on Discrepancies between Research and Practice," Paper presented at the Annual Meeting of the Academy of Management, 2007. See also M. London, E. M. Mone, and J. C. Scott, "Performance Management and Assessment: Methods for Improved Rater Accuracy and Employee Goal Setting," *Human Resource Management* 43 (2004), pp. 319–336.

functions: teaching, research, and service. Performance in each of these three areas is defined with different outcome measures. Students are obviously one source of data to evaluate the quality of the teaching. Performance in this context would involve outcome measures that define the "quality" of performance.

Performance on the job as a whole would be equal to the sum (or average) of performance on the major job functions or activities. For example, the World Bank identified eight job functions for its trainers (e.g., use of relevant examples, participant involvement, evaluation procedures). The functions have to do with the work that is performed and *not* the characteristics of the person performing. Unfortunately, many performance appraisal systems confuse measures of performance with the traits, or competencies, of the person.

Recommendation: Don't confuse performance with competencies

Let us emphasize this again: The definition of performance refers to a set of outcomes produced during a certain period and does *not* refer to the traits, personal characteristics, or competencies of the performer. (See Critical Thinking Application 7-A.) There is clearly a place for the assessment of competencies, knowledge, skills, and other personal characteristics of the performer. There is also a critical place for an assessment of behaviors on the job but these behaviors should be defined and ultimately assessed in terms of desirable or undesirable outcomes that may derive from these behaviors. Our main point here is that **there should be a clear distinction between the measurement of the person and his or her skills, knowledge, competencies, or potentiality and that person's actual performance.** Such factors are surely correlated with performance outcomes but they are not the same thing as performance. Their measurement should thus be viewed as diagnostic in the context of performance appraisal and should be used to assess the **potential** to perform and to (hopefully) improve the record of performance outcomes. But diagnostic assessments or judgments of potential and measures of exhibited performance are very different things.

Pick any sport to underscore this distinction. A golfer records an 18-hole score. This is one simple measure of her performance (we could also break her performance down into much more precise elements of that performance such as the number of putts, drives in the fairway, sand saves, etc.). A breakdown of her swing or her putting stroke would be a diagnostic assessment that could be made in an effort to improve a particular performance measure.

The most effective PM&A systems define and measure performance as clearly as possible in the context of carefully defined organizational objectives and then attempt to understand the causes of that good or not so good performance. It's clearly more difficult to draw this distinction for most jobs outside of sports but it can still be done and is done. One objective we have in this chapter is that you will understand how this can be done by the time you have finished reading the chapter. And, yes, this objective is a simple example of a performance objective that could ultimately be measured with an outcome (e.g., did you understand how performance should be defined?). An appropriate performance measure for this objective could be something like "I have a clear and unambiguous understanding of the difference between a measure of performance and a measure of some correlate of that performance." As the writers of this chapter, our goal is that 100 percent of the readers (our customers) would indicate that they do have this level of understanding.

What Are the Uses of Performance Data?

The information collected from PM&A systems is typically used for compensation, performance improvement or management (e.g., personnel decision making), and documentation. As discussed in Chapter 6, performance data are often used for staffing decisions (e.g., promotion, transfer, discharge, terminations, layoffs), and this is where the entire PM&A system may fall under the close scrutiny of the courts. PA is also used for training needs analysis, employee development, and research and program evaluation (e.g., validation research for selection methods).

Performance Management and Compensation

Performance appraisal information is often used by supervisors to manage the performance of their employees. PM&A data can reveal employees' performance weaknesses, which managers can refer to when setting goals or target levels for improvements. A PM&A system should include a diagnostic component where an evaluator attempts to explain a performance level or outcome based on a performer's behaviors, traits, competencies,

abilities, or motivations. But an effective system should first measure the performance level as accurately as possible and then attempt to explain the obtained level based on a performer's characteristics (competencies, KASOCs). One of the strongest trends in this country is toward some form of pay-for-performance (PFP) system. Chapter 11 covers the important area of PFP, a critical component for effective compensation and, as evidenced by the recent economic meltdown, an HR function with the potential to also do great harm to an organization.

Internal Staffing

Performance appraisal information is also used to make staffing decisions. As discussed in Chapters 5 and 6, many organizations rely on performance appraisal data to decide which employees to move upward (promote) to fill openings and which employees to retain as a part of "rightsizing" (or downsizing) efforts. Performance appraisals should also be the basis of terminations when the organization concludes that performance fails to meet a minimum or acceptable standard or that, perhaps, the organization could do better without an employee (or with an alternative employee or work source).

One problem with relying on performance appraisal information to make decisions about job movements is that employee performance is typically measured only for the *current* job. If the job at the higher, lateral, or lower level is different from the employee's current job, then it may be difficult to estimate how the employee will perform on the new job if that new job requires significantly different competencies (or KASOCs). Assessments of these competencies can be done in a variety of ways, including judgments by supervisors, peers, and even subordinates. Of course, many organizations use assessment tools such as those described in Chapter 6.

Assessments of competencies or other worker characteristics using ratings by qualified rating sources such as supervisors and peers are a perfectly acceptable approach for internal staffing decisions and, in many cases, more valid than other approaches to assessment, such as those discussed in Chapter 6. However, such assessments should be distinguished from the measurement of performance.

"Predictive weights" for PA data

It is possible to apply "predictive weights" to performance appraisal data to use the data for promotional decisions. If a study establishes a linkage between effective performance on certain job dimensions of Job A with effective performance in Job B, then ratings on those dimensions for Job A performance could be given predictive weights depending on their relative ability to predict performance. But it is not advisable to rely only on performance appraisal data to make promotional decisions since the jobs are undoubtedly different to an extent and thus may require somewhat different KASOCs or competencies.[13] Of course, the extent of these differences is related to the predictive value of the performance measurement. Performance in a sales job may or may not be related to performance as the sales manager. Performance as a retail assistant manager may be highly predictive of performance of the store manager.

Training Needs Analysis

Most firms use appraisal data to determine employees' needs for training or development. Hundreds of companies, including Microsoft, IBM, and Merck, now use 360-degree or multisource appraisal (e.g., subordinates, peers, clients) as feedback for their supervisors or managers.[14] The results are revealed to each manager with suggestions for specific training and development (if needed). Honeywell, for example, has specific training modules based on 360-degree appraisal ratings on several job functions.

Many organizations have adopted social networking methodologies to improve performance feedback. Accenture has a Facebook-style program called Performance Multiplier where employees can post work status updates, photos, and goals that can be viewed by fellow staffers. Rypple lets people post Twitter-length questions about their performance in exchange for anonymous feedback. These questions can go out to clients, peers, subordinates, and managers. Among the companies using the Rypple software are Harvest Bread Co. and Mozilla.[15]

Research and Evaluation

Performance data can also be used to determine whether various human resource programs (e.g., particular selection methods, training programs, recruitment sources) are effective.[16] For example, when the City of Toledo, Ohio, wanted to know whether its police officer

selection test was valid and job related, it collected performance appraisal data on officers who had taken the test when they were hired so that test scores could then be correlated with job performance ratings. We know that better and more comprehensive measures of performance can provide stronger (and more legally defendable) evidence for establishing the "job relatedness" of selection methods.[17]

LEGAL ISSUES ASSOCIATED WITH PERFORMANCE APPRAISALS

Since performance appraisal data are often used to make many important personnel decisions (e.g., pay, promotion, selection, termination), it is understandable that appraisal is a major target of legal disputes involving employee charges of unfairness and bias.[18] There are several legal avenues a person may pursue to obtain relief from discriminatory performance appraisals. As discussed in Chapter 3, the most widely used federal laws are Title VII of the Civil Rights Act and the Age Discrimination in Employment Act. However, there are numerous other possible sources of redress.

There are several recommendations to assist employers in conducting fair performance appraisals and avoiding legal suits. Figure 7-3 presents a summary of these recommendations based on a recent study and reviews of court cases related to appraisal.[19] The figure lists 15 PA characteristics related to the content, process, and results of PA. They are presented in their approximate order of importance in the prediction of the outcomes of court cases involving PA. For example, a violation of the 80 percent rule using PA data to make personnel decisions was found to be the most important predictor of the outcome of cases such that a violation increases the probability that the plaintiff (or protected class of plaintiffs) would prevail in the lawsuit. Many allegations of discrimination in EEO cases involving performance appraisal focus on the level of "subjectivity" in the PA process. For example, expert testimony on behalf of the plaintiffs in several gender discrimination lawsuits emphasized the **"excessive subjectivity"** of the performance appraisal process where statistical prima facie evidence of discrimination was presented, and very few (if any) of the prescriptions in Figure 7-3 characterized these PA systems.[20]

15 Predictors of the outcomes of court cases

Recall the discussion in Chapter 3 about adverse impact related to personnel decisions and court rulings regarding the **"disparate impact"** theory of discrimination and performance appraisal. The Supreme Court has ruled that adverse impact statistics such as the 80 percent rule can be used in Title VII and ADEA cases where performance appraisal was used to make decisions regarding who gets promoted, who gets terminated (consider Ford's age and race discrimination case related to its downsizing; see Critical Thinking Application 7-C), who gets merit raises, and any other important personnel decisions.

80 percent rule can be used in PA cases

Organizations should audit their appraisal data to test for possible adverse impact effects long before they get sued. They might even avoid getting sued. Adverse impact statistics have also been used successfully in **"disparate treatment"** cases to support an individual's claim of race or gender discrimination. Plaintiffs have used such data to augment claims of "disparate treatment" discrimination indicating a "pattern or practice" of discrimination and to buttress a motion for "class certification" that resulted from the "extreme subjectivity" of a bad performance appraisal system.

Such data can be used by the employer to rebut such a claim if in fact there is no evidence of adverse impact related to a particular protected class. Bottom line for organizations: An organization is in trouble if it gets sued, and there is a certified class of alleged victims (e.g., a class of females, minority, or older workers), and the organization has violated the 80 percent rule in its decisions (e.g., promotions, terminations) based on the use of a flawed appraisal system that adheres to few (or none) of the recommendations in Figure 7-3. **Prima facie evidence such as the 80 percent rule is considered to be the single best predictor of the outcome of cases involving PA.**

80 percent rule violations–best predictor of case outcomes

Figure 7-3 **Employer Prescriptions for Winning Legal Challenges Regarding Performance Appraisal***

DID THE EMPLOYER:

1. Audit personnel decisions stemming from PA data to make certain there is not prima facie evidence of discrimination (e.g., 80 percent rule violations)?
2. Use procedures for performance appraisal that do not differ as a function of the race, sex, national origin, religion, disability, or age of those affected by such decisions?
3. Use objective or countable (nonrated) performance outcome data?
4. Have a formal system of review and appeal for situations in which the rated individual disagrees with a rating?
5. Use more than one independent evaluator of performance?
6. Use a formal, standardized system for the personnel decision?
7. Document that relevant evaluators have had ample opportunity to observe rated performance or to review work products (if ratings must be made)?
8. Rate behavior or outcomes and avoid ratings on traits such as dependability, judgment, drive, flexibility, aptitude, innovativeness, or attitude?
9. Validate/corroborate the performance appraisal data with other data?
10. Communicate precise and specific performance standards to employee?
11. Provide written instructions to raters on how to complete the performance evaluations?
12. Evaluate employees on specific work dimensions rather than a single overall or global measure of performance or promotability?
13. Require a consistent policy of documentation for extreme ratings (e.g., critical incidents)?
14. Provide employees with an opportunity to review their appraisals?
15. Train personnel decision makers on performance appraisal, rating errors, and laws regarding discrimination?

*These prescriptions are in their approximate order of predictive importance. Thus, assuming no 80 percent rule violations (item #1), employers with PA systems that meet these prescriptions are more likely to prevail in court challenges.

Source: H. J. Bernardin, "Legal Prescriptions Based on Expert Judgments of Performance Appraisal System Characteristics." Under review, Human Resource Management.

DESIGNING AN APPRAISAL SYSTEM

The process of designing an appraisal system should involve managers, employees, HR professionals, and, most important, internal and external customers in making decisions about each of the following issues.

- Measurement content.
- Measurement process.
- Control of rating errors and biases.
- Defining the rater (i.e., who should rate performance).
- Defining the ratee (i.e., individual, unit, organization).

It is a challenge to make the correct decisions since no single set of choices is optimal in all situations. The starting point should be the strategic plan and objectives of the organization. The details of the plan should be reviewed in order to design an appraisal system consistent with the overall goals of the firm. This is particularly true with regard to measurement content and the outcomes to be emphasized.

Measurement Content

As we discussed earlier, performance appraisal in practice is too often person-oriented and focused on a person's characteristics. PM&A systems should first be work-oriented and focus on the **record of outcomes** that the person achieved on the job. Effective *performance* appraisal focuses on the record of outcomes and, in particular, outcomes directly linked to an organization's mission and objectives.[21] Some Sheraton Hotels offer 25-minute room service or the meal is free. Sheraton employees who are directly connected to room service are appraised on the record of outcomes specifically related to this service guarantee. Lenscrafters guarantees new glasses in 60 minutes or they're free. Individual and unit performance are measured by the average time taken to get the new glasses in the customer's hands. These are outcomes. *In general, personal traits*

Traits or competencies are correlates of performance–not performance

or characteristics (e.g., dependability, integrity, motivation, perseverance, knowledge, attitude, loyalty) should not be used when evaluating past performance since these constructs are not measures of actual performance. As personal characteristics of a performer, they may very well be correlates or predictors of performance, but they are *not* measures of actual performance. They should be assessed but not as surrogate measures of performance.

Performance can be (and usually should be) defined in terms of both countable quantitative output (or outcome) measures and also by ratings made by supervisors, customers, and others. Some examples of countable results are units produced or sold, sales, the number of customers served, error rates, breakage or waste, the number of publications, and grant proposals submitted or funded. While the number of these measures is tabulated and not rated, the effectiveness of a particular level of output is typically rated by someone. College professors at research-oriented universities often have performance objectives such as publishing research in leading academic journals. First, someone must define a "leading academic journal." Then, a measure of performance for the professor on her research activities could be the "x" number of publications in these journals for a given period. This record of outcomes must still then be rated for effectiveness (e.g., does one publication in a year constitute effective or ineffective performance?). This rating of effectiveness may determine the professor's tenure or whether she gets a raise for her level of performance.

President Obama's $4.3 billion education initiative, known as the "Race to the Top," includes federal funding for what's called pay-for-performance.[22] This is one reason many states are moving toward performance measurement systems for teachers that include much more emphasis on objective data such as student test scores. A strong trend is the requirement that teacher pay systems place less emphasis on rewarding college degrees held and years on the job and more emphasis on how much students learn. The theory, of course, is that teachers will work harder if they know their pay (and perhaps their continued employment) depends on how their students perform.

There is already much greater reliance on the use of student test scores to define and measure teacher performance. Based on a 2011 law, Florida teachers are now assessed by a new test-based evaluation system and could lose their jobs for poor performance based on students' test performance. The state is also developing a "value-added" system to judge teacher quality with test-score data that would take into account those factors (constraints) that are outside of a teacher's control.[23]

Categories of Performance Outcomes

There are six categories of performance outcomes by which the value of performance in any work activity or work function may be assessed.[24] These six criteria are listed and defined in Figure 7-4. Although all of these criteria may not be relevant to every job activity or job function, a subset of them will be. It is also important for organizations to recognize the relationships among the criteria. For example, sometimes managers encourage employees to push for quantity, without recognizing that quality may suffer or that

Figure 7-4 **The Six Primary Criteria on Which the Value of Performance May Be Assessed**

1. *Quality:* The degree to which the process or result of carrying out an activity approaches perfection, in terms of either conforming to some ideal way of performing the activity or fulfilling the activity's intended purpose.
2. *Quantity:* The amount produced, expressed in such terms as dollar value, number of units, or number of completed activity cycles.
3. *Timeliness:* The degree to which an activity is completed, or a result produced, at the earliest time desirable from the standpoints of both coordinating with the outputs of others and maximizing the time available for other activities.
4. *Cost-effectiveness:* The degree to which the use of the organization's resources (e.g., human, monetary, technological, material) is maximized in the sense of getting the highest gain or reduction in loss from each unit or instance of use of a resource.
5. *Need for supervision:* The degree to which a performer can carry out a job function without either having to request supervisory assistance or requiring supervisory intervention to prevent an adverse outcome.
6. *Interpersonal impact/contextual or citizenship performance:* The degree to which a performer promotes feelings of self-esteem, goodwill, and cooperation among co-workers and subordinates.

co-workers might be affected. Likewise, they may focus on quantity without emphasizing timeliness, cost effectiveness, quality, or interpersonal impact. Emphasis on one particular outcome category (e.g., quantity) can obviously have an impact on some other category of outcomes, particularly quality.

Contextual performance

The interpersonal criterion includes **"contextual or citizenship performance"** as discussed in the literature.[25] A good "organizational citizen" is an employee who contributes beyond the formal role expectations of a job as might be detailed in a job description. Such employees are positively disposed to take on alternative job assignments, respond cheerfully to requests for assistance from others, are interpersonally tactful, arrive to work on time, and often may stay later than required to complete a task. Contextual performance operates to either support or inhibit technical production and can facilitate individual-, group-, and system-level outcomes.[26] As we discussed in Chapter 6, contextual performance can also be defined in terms of "workplace deviance" or counterproductive behaviors.[27]

Contextual performance contributions such as mentoring, facilitating a pleasant work environment, and compliance with organizational and subunit policies and procedures may have implications for several of the other outcome categories as well. If performance is defined at a more specific task or activity level, contextual performance also could be represented in the description of the function itself and combined with one or more of the value criteria (e.g., quality, quantity). For example, one model of "citizenship performance" includes "personal support" as a dimension and defines it by such behaviors as "helping others by offering suggestions, teaching useful knowledge or skills, and providing emotional support for their personal problems."[28] We could certainly define outcomes in these areas according to quantity and quality values (e.g., how often is emotional support offered; how good was it?).

Measuring Overall Performance

While an overall rating approach where the rater is not asked to distinguish among the criteria is surely faster than making assessments on separate criteria, the major drawback is that it requires raters to simultaneously consider perhaps as many as six different aspects of value and to mentally compute their average. The probable result of all this subjective reasoning may be less accurate ratings than those done on each relevant criterion for each job activity and less specific feedback to the performer. *In general, the greater the specificity and precision in the content of the appraisal, assuming the content is compatible with the strategic goals of the organization, the more effective the appraisal system regardless of the purpose for the appraisal system*[29] (see Figure 7-1 again).

The Measurement Process

There are three basic ways in which raters can make performance assessments: (1) they can make comparisons of ratees' performances, (2) they can make comparisons *among* anchors or standards and select one most descriptive of the person being appraised, and (3) they can make comparisons of individuals' performance *to* anchors or standards. These are shown in simplified form in Figure 7-5. Some of the most popular or promising rating instruments representing each of these three ways are described next.

Rating Instruments: Comparisons among Ratees' Performances

Paired comparisons, straight ranking, and forced distribution are appraisal systems that require raters to make comparisons among ratees according to some measure of effectiveness or simply overall effectiveness. Although controversial, employee comparison systems are growing in popularity to some extent because Jack Welch, GE's famous retired CEO, has been a strong advocate of the approach for many years.

Paired comparisons require the rater to compare all possible pairs of ratees on "overall performance" or some other, usually vaguely defined, standard. This task can become cumbersome for the rater as the number of employees increases and more comparisons are needed. The formula for the number of possible pairs of employees is $n(n-1)/2$, where n is the number of employees. **Straight ranking,** or rank ordering, asks the rater to simply identify the "best" employee, the "second best," and so forth, until the rater has identified the worst employee. For example, some NCAA rankings in football and basketball are based on a rank ordering of the teams by coaches and the press. Ranking systems are popular in research labs such as Sandia and Lawrence Livermore. Managers are forced to rank their subordinates in a 1 to N order based on performance.

Figure 7-5
Rating Format Options

COMPARISONS AMONG PERFORMANCES

Compare the performances of all ratees to each anchor (or standard) for each job activity, function, or overall performance. Rater judgments may be made in one of the following ways:

- Indicate which ratee in each possible pair of ratees performed closest to the performance level described by the anchor or attained the highest level of overall performance. (Illustrative method: paired comparison)
- Indicate how the ratees ranked in terms of closeness to the performance level described by the anchor or standard. (Illustrative method: straight ranking)
- Identify a predetermined percentage of employees as ineffective and highly effective. (Illustrative method: forced distribution)

COMPARISONS AMONG ANCHORS

Compare all the anchors for each job activity or function and select the one (or more) that best describes the ratee's performance level. Rater judgments are made in the following way:

- Indicate which of the anchors fit the ratee's performance best (and/or worst). (Illustrative method: Computerized Adaptive Rating Scales, forced choice)

COMPARISONS TO ANCHORS

Compare each ratee's performance to each anchor for each job activity or function. Rater judgments are made in one of the following ways:

- Whether or not the ratee's performance matches the anchor. (Illustrative methods: graphic rating scales such as Behaviorally Anchored Rating Scales; Management By Objectives)
- The frequency with which the ratee's performance matches the anchor. (Illustrative methods: all summated rating scales such as Behavioral ObservationScales and Performance Distribution Assessment)
- Whether the ratee's performance was better than, equal to, or worse than that described by the anchor. (Illustrative method: Mixed Standard Scales)

Forced distribution usually presents the rater with a limited number of categories (usually three to seven) and requires (or "forces") the rater to place a designated portion of the ratees into each category. A forced distribution usually places the majority of employees in the middle category (i.e., with average ratings or raises) while fewer employees are placed in higher and lower categories.

Some organizations use forced distribution to ensure that raters do not assign all (or nearly all) of their employees the most extreme (e.g., highest) possible ratings. Ford adopted a forced letter grade system for each supervisor. Thus, only 10 percent of employees could receive an A grade while the first 10 percent (and later 5 percent) had to receive a C grade. Employees who received Cs were not eligible for a raise or a bonus, and two C grades in a row could result in demotion and termination.

In addressing GE shareholders former CEO Jack Welch stated, "A company that bets its future on its people must remove that lower 10 percent, and keep removing it every year—always raising the bar of performance and increasing the quality of its leadership."[30]

Research on forced discrimination is not favorable

Research on forced distribution is not favorable which may explain why GE no longer relies on it. Enron had a forced-distribution system in place when the company collapsed and Microsoft dumped its system in 2008. Companies using forced distribution found an improved range in performance ratings, a primary purpose of the approach, but a lower overall evaluation of the appraisal system compared to other approaches. Supervisors and managers are often offended that no matter how effective they are as managers they must comply with the required forced distribution.[31]

Rating Instruments: Comparisons among Performance-Level Anchors

Computerized adaptive rating scales (CARS) is a promising rating method that presents raters with pairs of behavioral statements reflecting different levels of performance on the same performance dimension.[32] For example, for the performance dimension entitled "Personal Support," raters could be asked to select one of the following two statements as most descriptive of a particular ratee.

1. Refuses to take the time to help others when they ask for assistance with work-related problems.
2. Occasionally takes the time to help others when they ask for assistance with work-related problems.

Based on the statement selected, additional statements are then paired through a computer program for subsequent rating. The new pair of behavioral statements would then be selected, one of which was scaled by experts to be somewhat higher in effectiveness than the one first selected and the other of which was scaled to be somewhat below the level of effectiveness represented by the first statement selected. A rater's selection from this next pair of statements would then revise the estimate of the ratee's performance effectiveness level. Based on this new estimate, two new statements are selected by the computer program until the performance level can be measured reliably. However, there is no field research with CARS.

Laboratory research with CARS supported this method when compared to behaviorally anchored rating scales (BARS) and simple graphic rating scales.

Forced choice designed to reduce intentional bias

Forced choice is another PA method that requires the rater to compare performance statements and select one (or more) as most descriptive. Unlike the CARS method, the forced choice method is specifically designed to reduce (or eliminate) intentional rating bias where the rater deliberately attempts to rate individuals high (or low) irrespective of their performance. The rationale underlying forced choice is that statements are grouped in such a way that the scoring key is not known to the rater (i.e., the way to rate higher or lower is not apparent). The rater is unaware of which statements (if selected) will result in higher (or lower) ratings for the ratee because all statements appear equally desirable or undesirable. For example, if you were asked to select the two statements that are most descriptive of your instructor for this class, which two would you select?

1. Is patient with slow learners.
2. Lectures with confidence.
3. Keeps the interest and attention of the class.
4. Presents objectives before each class session.

The statements are chosen to be equal in desirability in order to make it more difficult for the rater to pick out the ones that will give the ratee the highest or lowest ratings based on personal bias. However, only two of the items actually distinguish highly effective from ineffective performers. For the present case, items 1 and 3 have been shown to discriminate between the most effective and the least effective college professors. Items 2 and 4 did not generally discriminate between effective and ineffective performers. If you selected statements 1 and 3 as most descriptive of your instructor, then he or she would be awarded two points. This procedure would be used with each of the 20 to 40 groups of statements to determine the total score for each ratee. Raters are not given the scoring scheme, so they are unable to intentionally give performers high or low ratings. Research with forced choice is limited, but there is some evidence that deliberate bias can be reduced with this method. Unfortunately, raters do not like this method specifically because the scoring key is hidden and they are often surprised by results.[33]

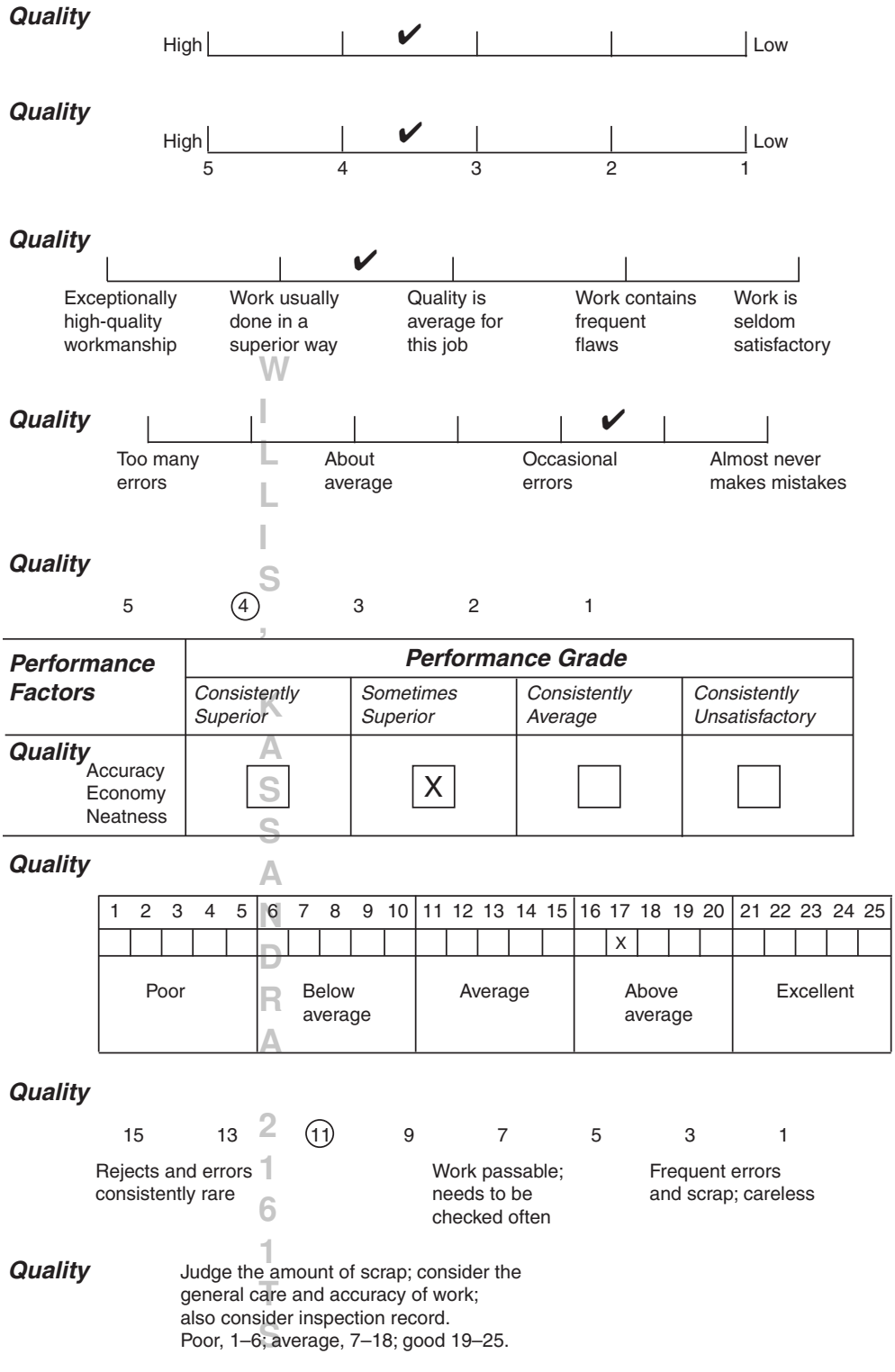
Deliberate bias can be reduced with forced choice

Rating Instruments: Comparisons to Performance-Level Anchors

Methods that require the rater to make comparisons of the employee's performance to specified anchors or standards include graphic rating scales, behaviorally anchored rating scales (BARS), management by objectives (MBO), summated scales (e.g., behavioral) observation scales (BOS), and performance distribution assessment (PDA). *Graphic rating scales* are the most widely used type of rating format. Figure 7-6 presents some examples of graphic scales. Generally, graphic rating scales use adjectives or numbers as anchors, but the descriptive detail of the anchors differs widely.

One of the most heavily researched types of graphic scales is **behaviorally anchored rating scales (BARS).** As shown in Figure 7-7, BARS are graphic scales with specific behavioral descriptions defining various points along the scale for each dimension. The recommended rating method for BARS asks raters to record specific observations or critical incidents of the employee's performance relevant to the dimension on the scale.[34] In Figure 7-7, the rater has written in "Stuck to the course outline, . . ." between points 9 and 10 on the left side of the scale. The rater would then select that point along the right side of the scale that best represents the ratee's overall performance on that function. That point is selected by comparing the ratee's actual observed performances to the behavioral

Figure 7-6
Examples of Graphic Rating Scales

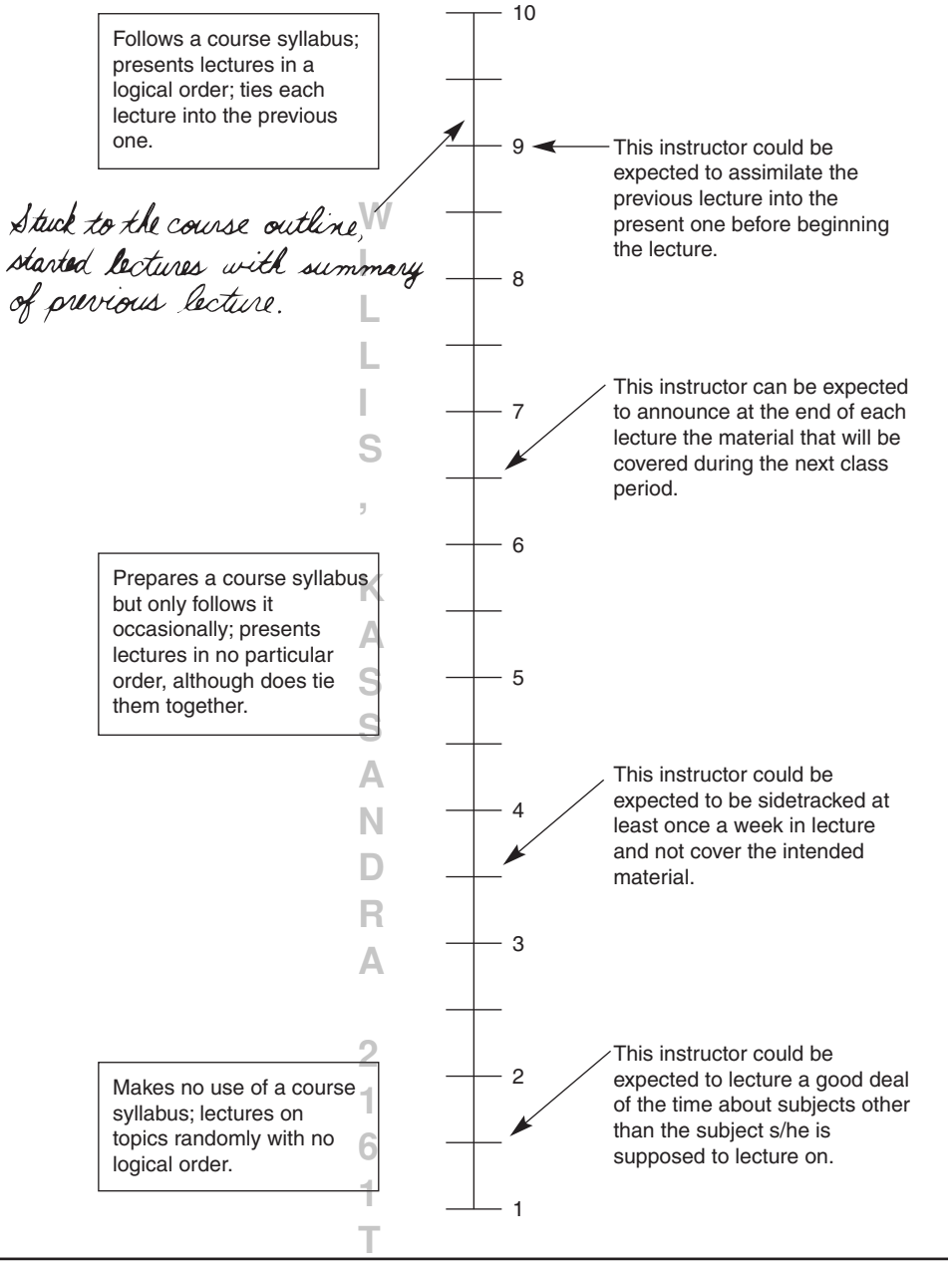


Source: R. M. Guion, *Personnel Testing,* 1965, p. 98. Reprinted with permission.

expectations that are provided as "anchors" on the scale. The rationale behind writing in observations on the scale prior to selecting an overall anchor point is to ensure that raters are basing their ratings of expectations on actual observations of performance. In addition, the observations can be given to ratees as feedback on their performance. Research shows that this form of feedback, with all recorded observations or critical incidents, is effective

Figure 7-7
An Example of a Behaviorally Anchored Rating Scale

Organizational skills: A good constructional order of material slides smoothly from one topic to another; design of course optimizes interest; students can easily follow organizational strategy; course outline followed.



Source: From Bernardin/Beatty. Performance Appraisal, 1e. © 1984 South-Western, a part of Cengage Learning, Inc. Reproduced by permission. www.cengage.com/permissions

BARS method improves performance

at improving performance and that this particular BARS approach is more effective than other formats at improving performance.[35]

The method of **summated scales** is one of the oldest formats and remains one of the most popular for the appraisal of job performance. One version of summated scales is **behavioral observation scales (or BOS).**[36] An example of a summated scale is presented in Figure 7-8. For this scale, the rater is asked to indicate how frequently the ratee has

Figure 7-8
A Summated Rating Scale

Directions: Rate your manager on the way he or she has conducted performance appraisal interviews. Use the following scale to make your ratings:

1 Always
2 Often
3 Occasionally
4 Seldom
5 Never

1. Effectively used information about the subordinate in the discussion.
2. Skillfully guided the discussion through the problem areas.
3. Maintained control over the interview.
4. Appeared to be prepared for the interview.
5. Let the subordinate control the interview.
6. Adhered to a discussion about the subordinate's problems.
7. Seemed concerned about the subordinate's perspective of the problems.
8. Probed deeply into sensitive areas in order to gain sufficient knowledge.
9. Made the subordinate feel comfortable during discussions of sensitive topics.
10. Projected sincerity during the interview.
11. Maintained the appropriate climate for an appraisal interview.
12. Displayed insensitivity to the subordinate's problems.
13. Displayed an organized approach to the interview.
14. Asked the appropriate questions.
15. Failed to follow up with questions when they appeared to be necessary.
16. Asked general questions about the subordinate's problems.
17. Asked only superficial questions that failed to confront the issues.
18. Displayed considerable interest in the subordinate's professional growth.
19. Provided general suggestions to aid in the subordinate's professional growth.
20. Provided poor advice regarding the subordinate's growth.
21. Made specific suggestions for helping the subordinate develop professionally.
22. Remained calm during the subordinate's outbursts.
23. Responded to the subordinate's outbursts in a rational manner.
24. Appeared to be defensive in reaction to the subordinate's complaints.
25. Backed down inappropriately when confronted.
26. Made realistic commitments to help the subordinate get along better with others.
27. Seemed unconcerned about the subordinate's problems.
28. Provided poor advice about the subordinate's relationships with others.
29. Provided good advice about resolving conflict.
30. When discussing the subordinate's future with the company, encouraged him/her to stay on.
31. Used appropriate compliments regarding the subordinate's technical expertise.
32. Motivated the subordinate to perform the job well.
33. Seemed to ignore the subordinate's excellent performance record.
34. Made inappropriate ultimatums to the subordinate about improving performance.

Research on BOS is positive

performed each of the listed behaviors. The ratings are then averaged or totaled for each person rated. Research on BOS is also quite positive,[37] especially when workers participate in the development of the scales.[38]

PDA measures constraints

Performance Distribution Assessment (PDA) is a more complicated rating method based on the theory that ratings should be made in the context of opportunities to perform at a certain level.[39] **PDA is the only method that statistically incorporates constraints on performance as a formal part of the measurement process.** For example, in evaluating managers on the quality of their performance "monitoring" at Tiffany's of New York, the managers are asked to consider how many opportunities the manager had to "furnish information in response to an inquiry that was completely accurate with respect to central questions posed by the inquiry" and then to rate how frequently the manager achieved that level of performance. Although raters report some difficulty with the rating process, PDA provides detailed documentation of constraints on performance and thus has the potential to remove those constraints over time.[40]

Management by objectives (MBO) is a performance management and appraisal system that calls for a comparison between specific, quantifiable target goals and the actual results achieved by an employee. Although there are no recent surveys on this subject, MBO is still probably the most popular PM&A method for managers.[41] MBO is similar in many respects to the "Productivity Measurement and Enhancement System" (discussed later) but which is typically focused on unit and not individual performance.[42]

With MBO, the measurable, quantitative objectives or goals are usually mutually agreed upon by the employee and supervisor at the beginning of an appraisal period. During the review period, progress toward the goals is monitored. At the end of the review period, the employee and supervisor meet to evaluate whether the goals were achieved and to decide on new goals. The goals or objectives are usually set for individual employees or units and usually differ across employees (or units) depending on the circumstances of the job. For this reason, MBO has been shown to be useful for defining "individual" or unit performance in the context of strategic plans. As a motivational technique, as long as the objectives that are set are defined in specific terms using carefully defined criteria as listed in Figure 7-4, attainable as perceived by the performer while still being difficult, MBO is an effective approach to improving performance and motivating employees. Thus, precise definitions of quality and quantity, specifically linked to unit strategic goals, can make MBO a very effective PM system. But the criteria must be defined (and ultimately evaluated) with strategic goals in mind.

MBO: Effective for improving performance

MBO is most often linked to countable results or performance indicators that may be rated for effectiveness by management but are counted objectively. This standard for effectiveness should have been set prior to the start of the appraisal period so no actual rating is necessary once the outcome count is counted. For example, the number of grant proposals submitted for a nonprofit grant writer would be the countable result that could be the focus of the MBO program. However, a determination of an acceptable level of effectiveness for this countable result would be made by management with participation and consultation with the performers. MBO is more effective when the attainable objectives are defined prior to the appraisal period in terms of nonrated or countable results.

MBO is not recommended as a method for comparing people or units unless the objectives that are set can be judged to be equally attainable in the context of potential situational constraints on performance, as discussed in a later section.

What Is the "Bottom Line" on What It Is We Should Be Measuring?

You'll note if you look back at Figure 7-1 that one of the recommendations nested under "strive for as much precision in defining and measuring performance dimensions as is feasible" is to use important objective data wherever possible but to appraise performance using ratings of **relative frequencies** when ratings are necessary. There are many options for rating behavior or performance outcomes. First of all, as stated earlier, performance appraisal should focus on the *record of outcomes.* Let's pick on your instructor for a bit. Certainly "instruction" is a critical activity (i.e., function) of his or her job. You are a type of "customer" who should be evaluating performance on that function (note prescription no. 2 in Figure 7-1 too). The most important criterion to define the "quality" of instruction is probably how much you actually learn from the instruction. Perfection (i.e., the highest level of effectiveness) would then be a perfect score on some test of the knowledge you were supposed to acquire from this instruction. For a number of reasons, scores on such a test may not be a practical source of data and such data, with some exceptions, probably does not allow for comparisons on instructors for decision-making purposes. Another way of getting at the "quality" of instruction is to have you (the customer) define levels of performance and then have you rate the extent to which the instructor meets or exceeds these levels of performance. Research says that when you do these ratings, your focus should be on how *frequently* (e.g., sometimes, 100 percent of the time, never) the instructor achieved this level of performance in the context of all the times the instructor had the opportunity to achieve at this level. Ratings of frequency are better than ratings of "intensity" (e.g., strongly agree/disagree) or satisfactoriness (e.g., how satisfied you are with your instructor), primarily because those who are rated regard "frequency" feedback as more accurate and helpful.

Recommendation: Rate on relative Frequency

Here's an example of "perfection" as defined by students to evaluate "instruction": "Every time a lecture was given, I had a clear, unambiguous understanding of what it was s/he was trying to teach; no questions were needed to clarify the material presented." Let's say this defines the "perfection" level of performance for "instruction." Raters would make ratings of relative frequency on the "quality of instruction" dimension by rating how often the instructor hit this level of performance out of all the times s/he gave lectures (or did instruction). So, you might give your instructor 100 percent on this dimension level; that is, every time a lecture was given, you had a "clear, unambiguous understanding." Obviously, it is also possible that the rating here could be 0 percent! That's why we also need to define at least one other level of performance for "quality of instruction." Raters would then rate how frequently the instructor achieved this level of performance. Research on rating formats shows that ratings of relative frequency result in higher levels of reliability in ratings (across raters rating the same person) and that the people who receive feedback on their performance actually prefer frequency ratings to other feedback options. The PDA system is most compatible with this approach to appraisal although BOS also calls for frequency ratings.

Higher reliability for ratings of relative frequency

Whatever is measured should obviously be vital to the strategic goals of the organization. Formal PM&A should clearly concentrate on reliable and valid measurement of outcomes that are directly linked to strategic goals and outcomes. One would hope that a strategic goal of your institution is superior instruction in every class. Sometimes the linkage between individual performance appraisal and the strategic objectives of the organization is very clear (a 100 percent score on "instruction" would be nice). In general, **the more precision in measurement, the better the PM&A system regardless of its purpose.** In general performance data that are compiled and not rated and that are clearly aligned with the strategic goals of the organization will be more effective performance measures than rated performance measures.

Recommendation: Tie performance measures to job descriptions

At the individual performer level, having performance dimensions that are derived from that performer's job description (or actually a part of it) makes a whole lot of sense. While some supervisors tend to ignore job descriptions, workers tend to look at job descriptions as contracts (i.e., this is why you're paying me).[43]

Control of Rating Errors and Biases

Performance observations and subsequent ratings are subject to a wide variety of inaccuracies and biases, often referred to as rating errors or rating effects.[44] These errors or biases occur during an observational period or when rater judgments must be made and can seriously affect performance appraisal results and validity. Unfortunately, many of these biases and errors cannot be eradicated easily (if at all).

There is a plethora of research on what can (and does) go wrong with PA. One expert on the subject referred to appraisal as the "Achilles heel" of human resource management.[45] While there are many reasons for the problems related to appraisal, the process itself is quite difficult even for those who fully intend to rate as accurately as they can. There are many unintentional errors in observation and rating. But then, of course, you have all sorts of other rater goals that are not necessarily related to accuracy at all.[46] When raters, usually supervisors, are judging their subordinates' performances, the political implications of their actions are often "front and center" in the consideration of the ratings to be given.[47] Appraisals often take place in organizations that are not necessarily operating in a rational manner. In such an environment, rating performance in an accurate manner is sometimes less significant to raters than other motives for rating at a certain level. Raters may be unwilling to rate accurately or may be uninterested in doing so.[48] Research also indicates that raters may be motivated to rate inaccurately. As one recent review put it, "the distinction here is whether raters have: (a) a lack of incentives to rate accurately, (b) ambivalence toward rating accurately, or (c) too many incentives to rate inaccurately."[49]

Some rater goals are not compatible with accuracy

There are clearly many "intentional" errors and biases. The major sources of biases and errors are discussed next. The most common rating errors are leniency/severity, halo/horns effect, central tendency, fundamental attribution errors, representativeness, availability, and anchoring.

Leniency/Severity

Leniency occurs when ratings for employees are generally at the high end of the rating scale regardless of the actual performance of ratees.[50] This error is usually considered to be deliberate but it has been linked to some rater personal characteristics. Surveys have identified leniency as the most serious problem with appraisals whenever they are linked to important decisions such as compensation, promotions, or terminations.[51] Research shows that leniency (or severity) in ratings is related to the personality and competence of the *rater*.[52] The Five-Factor Model discussed in Chapter 6 applies here. Raters who are low on "Conscientiousness" and high on "Agreeableness" tend to be more lenient. Raters high on "Conscientiousness" with moderate levels of Agreeableness are the most accurate raters.[53] Also, raters more focused on diagnosing and assessing performance tend to be less lenient.

Leniency related to raters' personalities

Leniency is the primary reason that companies have turned to forced distribution systems such as General Electric' "A, B, C" system where managers have to put a certain percentage of subordinates into the "C" (low performance) category. Jack Welch and others argue that differentiation of employees by performance and making certain that the most important positions in the organization (the "A" positions) are occupied by "A" players is a key contributor to competitive advantage. You do not want "C" or even "B" players in vital positions. Obviously, leniency error precludes an organization's ability to differentiate among employees and take action to reward the "A" players, move "C" performers out of key positions as soon as possible, and try to develop the "B" players into "A" players. While forced distribution eliminates leniency, it creates other serious problems.[54]

In general people do not like to evaluate other people and particularly dislike confrontations with those who are rated. This is one of the main reasons that leniency occurs. One study showed promise for alleviating leniency. **"Self-efficacy training for raters"** provided training in giving negative feedback and produced less lenient ratings than a control group. This training involved observing a successful rater, a simulated appraisal session with a "problem" employee, feedback on performance, and then coaching on how to conduct an appraisal discussion. In addition to the reduced leniency, the research showed that "self-efficacy" training resulted in more positive perceptions of procedural fairness, more agreement in ratings between raters and performers, and, most important, higher unit performance.[55]

Self-efficacy training reduces leniency

Halo/Horns Effect

"Halo or horns" effect occurs when a rating or impression of one dimension (or, more often an overall impression) of an employee influences the ratings on other dimensions for the same employee. That is, the rater inappropriately assesses ratee performance similarly across different job functions, projects criteria, or performance dimensions. This error is not deliberate. Research indicates that rater training and more precisely defined rating scales can control this error to an extent.[56]

Central Tendency

Central tendency occurs when ratings for employees tend to be toward the center (midpoint) of the scale regardless of the actual performance of ratees. This is a deliberate error although much less common (and problematic) than leniency.

Fundamental Attribution Error/Actor-Observer Bias

The fundamental attribution error refers to the tendency to attribute observed behaviors or outcomes to the disposition of the person being observed while underestimating the causal role of factors beyond the control of the performer.[57] This is related to the **actor-observer bias** where people tend to make the exact opposite attributions for their *own* behavior: they tend to attribute their successes to their own competence and their failures to the influence of external factors beyond their control. **The actor-observer bias is thus the tendency of observers to underestimate the effects of external factors and for performers to overestimate the effects of external factors on less than perfect performance.** Rating systems such as **PDA** that ask the rater to formally consider the possible constraints on performance have been shown to reduce the actor-observer bias and decrease differences between self and supervisory appraisals.[58]

Actor-observer bias is one of the major factors that cause perceptions of unfairness in appraisal decisions. Any student who has been graded on a group project may have experienced this problem in individual appraisal. Many conditions present in the job situation or work assignment can hold a person back from performing as well as he or she

could. Some of these constraints include inadequate tools, lack of supplies, not enough money, too little time, lack of information, breakdowns in equipment, ineffective management, and not enough help from others. For example, truck inspectors may be limited in the number of trucks they can check for defects if they spend a considerable portion of their workday in court presenting testimony against offenders. They still may be held accountable, however, for inspecting a certain number of trucks despite these other job duties. If in a group project, one of your team members fails to retrieve vital information, the constraint could seriously hamper your ability to do your tasks. Situational factors that hinder an employee's job performance are called **situational constraints** and are described in Figure 7-9.[59]

Rater training should focus on actor/observer bias

An appraisal system should consider the effects of situational constraints so that ratees are not unfairly downgraded for these uncontrollable factors. Rater training programs also should focus on making raters aware of potential constraints on employee performances and the tendency on the part of raters to commit this attributional error. Research shows that training on the actor-observer bias can reduce the error and promote more agreement between the rater (observer) and the ratee (the actor).[60]

A goal-based PA that documents constraints

Figure 7-10 presents an example of an MBO-type system that considers the potential for this common error. With this method, raters and ratees must independently complete a performance dimension/constraint matrix. This approach places the focus squarely on discrepancies in perceptions of the effects of particular constraints. In Figure 7-10, the list of constraints was recorded by the ratee (performer) who felt that the constraint had a significant impact on her performance for a particular performance dimension. For example, this head of a research and development unit indicated that staff attendance at meetings was an indication of poor subordinate performance and that this constraint impeded performance on "Organizing and Conducting Seminars." After the supervisor has reviewed the constraints and recorded his/her own assessment of the effects of the constraints, specific goals are set where the supervisor agrees to attend to some (or all) of the perceived constraints (e.g., supervisor will send out a memo strongly encouraging seminar attendance).

Figure 7-9
Possible Situational Constraints on Performance

1. Absenteeism or turnover of key personnel.
2. Slowness of procedures for action approval.
3. Inadequate clerical support.
4. Shortages of supplies and/or raw materials.
5. Excessive restrictions on operating expenses.
6. Inadequate physical working conditions.
7. Inability to hire needed staff.
8. Inadequate performance of co-workers or personnel in other units on whom an individual's work depends.
9. Inadequate performance of subordinates.
10. Inadequate performance of managers.
11. Inefficient or unclear organizational structure or reporting relationships.
12. Excessive reporting requirements and administrative paperwork.
13. Unpredictable workloads.
14. Excessive workloads.
15. Changes in administrative policies, procedures, and/or regulations.
16. Pressures from co-workers to limit an individual's performance.
17. Unpredictable changes or additions to the types of work assigned.
18. Lack of proper equipment.
19. Inadequate communication within the organization.
20. The quality of raw materials.
21. Economic conditions (e.g., interest rates, labor availability, and costs of basic goods and services).
22. Inadequate training.

Figure 7-10 A Performance/Constraint Matrix (R&D) Director

	Performance Dimensions			
Constraints	**Assisting Center Researchers**	**Generating Research Grants**	**Organizing and Conducting Seminars**	**Conducting Research**
Absenteeism/turnover		a		
Slow procedures				
Clerical support		b		b
Supply shortage				
Excessive restrictions	c	d		
Working conditions				
Poor co-worker performance				
Poor subordinate performance			e	f
Poor manager performance				
Inefficient structure				
Excessive reporting requirements				
Workloads				
Change in administrative policy				
Co-worker pressure				
Change in work assignment	k, h	k, h	k, h	k, h
Lack of equipment				
Inadequate communication				i
Raw material problem				
Economic conditions				j
Lack of (or poor) training				

Constraints	**Goals**
a. Loss of departmental secretary precluded proposal writing for two months	a. Proposed backup clerical support for excessive workloads; have plan by 3/1
b. Secretary worked on unrelated project for two months	b. Develop more detailed job description and chain of command for secretaries, that is, only one boss; submit plan by 3/1
c. Grant support lifted from four recipients due to lack of funds	c. Review committee will be made aware of total funds available
d. No money allotted for hiring grantsperson as promised	d. Conduct search to determine if part-time person can be identified; write announcement by 2/15
e. Staff rarely attended seminars although they were scheduled on payday	e. Send memo to staff encouraging attendance
f. Staff member failed to do literature review in a collaborative research project	
h. Given new responsibility for compensation policy and computer records (not on original job description)	h. Provide written charge in the future
i. Failure of management to provide written charge for compensation project resulted in time being wasted in clarification with divisions	i. Get commitment from management to attend all executive-level meetings
j. Severe reduction in research budget has precluded three pilot projects that had great potential for external funding	j. More active search for external dollars. Will review foundation interests; submit report by 4/1
k. Asked to conduct seminar at last moment due to funding problem; took 15 percent of my time away from all assignments	k. Will do survey to determine what time would be most favorable; report attendance to director (will submit report in two months)

Representativeness Error

This error refers to the tendency to make judgments about people (or their performance) on the basis of their similarity to people who exhibited prominent or memorable levels on the attribute being judged, even though the similarity may have no causal connection to the attribute. For example, popular stereotypes may have given the rater an image that attractive people are more effective in groups or that males are more effective managers.

When confronted with the task of rating someone on factors related to his or her group effectiveness, to the extent that the ratee possesses the *representative* trait (e.g., is a male or is attractive), the rater will tend to rate in accordance with this preconception rather than in accordance with actual observations.

The problem with this type of stereotyped thinking is that it ignores the fact that although some prominent examples of people who performed at the upper or lower extremes of effectiveness may have possessed a certain characteristic, such as attractiveness, most of the people who possess such a characteristic do not so distinguish themselves, and, in fact, the characteristic has no causal connection to actual performance at all.

This is a difficult tendency to overcome and is more likely to occur when there is less precision in a PM&A system and more "subjectivity" in the appraisal process.[61] Perhaps the best means of suppressing it is to use rating scales that are anchored with detailed descriptions of behaviors or outcomes and to train raters in the tendency.

Availability Bias

People tend to mistake the ease with which a category of outcomes can be recalled as an indication of its frequency of occurrence relative to other categories. This becomes almost a rule of thumb that some people use in judging the relative frequency of outcomes. The relevance to performance appraisal judgments should be obvious: since more extreme outcomes tend to be more memorable, raters will tend to attribute greater frequency to them than was actually the case. This results in such outcomes being given excessive weight in the formation of appraisal judgments. It has been found that negative events—instances of ineffective performance—seem to have the greatest availability in memory.

There is no easy solution to the availability problem. It is possible that merely making raters aware of their proneness to this type of error will cause them to make efforts to compensate for this tendency. However, there is no research to substantiate this possibility.

Anchoring Error

This error refers to the tendency to insufficiently alter a judgment away from some starting point when new information is received.[62] Most of us start with some initial impression of any situation we encounter, or we form one very quickly after our initial immersion in a situation. This is very true of observations of other people's behavior or performances. Either from past experience, stereotyping, information available, interpersonal affect, or reputation, we generally start off prior to observing another's performance with some initial impression, or we form one very quickly. The problem that arises is that once an initial starting point, or **anchor,** is selected, we tend to resist being moved from this point by subsequent information that warrants movement. As a consequence, our final judgments will be much nearer to our "starting point" than they should be. This is a source of unfairness in appraisals. A person's reputation, or even his/her past performance, should not be a factor in how his/her performance during the period under consideration is rated.

Anchoring can bias multi-rater systems

Anchoring is a potent error in judgments of all kinds. For example, if a person whom I regard as unreliable and untrustworthy told me that the performance of a new hire had been terrible on his/her last job, even though I had many other sources of credible and contradictory information, I could be affected by that person's opinion in evaluating the new hire and even in subsequent evaluations of the new hire. This **anchoring** effect also applies to multirater systems. Supervisors, for example, can be inappropriately affected by the level of subordinates' initial self-ratings, particularly if the supervisor has not anchored future judgments with his/her own prior judgments. Supervisors should make assessments before they review (and consider) self-ratings and also be wary of their own preconceived notions.[63] It is also possible to anchor your own ratings and bias subsequent ratings.

The origin of this problem again seems to be the holistic consideration of a person's performance on each rating factor rather than attending to the specific behaviors or outcomes that were exhibited. Again, training on this potential error may be helpful. If rating scales are used that don't call for an overall judgment but rather elicit estimates of the frequencies

with which the behaviors or outcomes anchoring each level occurred, we might overcome (or reduce) the problem of anchoring.

Rater Training

All of these rating errors and biases can arise in two different ways: as the result of *unintentional* errors in the way people observe, store, recall, and report events or as the result of *intentional* efforts to assign inaccurate ratings. If rating errors are *unintentional,* raters may commit them because they do not have the necessary knowledge or skills to make accurate observations and ratings, or perhaps the criteria for the appraisal are not carefully defined. Rater training can help.

Attempts to control unconscious, unintentional errors most often focus on rater training. Training to improve a rater's observational, categorization and rating skills is called **frame-of-reference training (FOR)**.[64] This training consists of creating a common frame of reference among raters in the observation process. Raters are familiarized with the rating scales and are given opportunities to practice making ratings. Following this, they are given feedback on their practice ratings. They are also given descriptions of critical incidents of performance that illustrate outstanding, average, and unsatisfactory levels of performance on each dimension. This is done so they will know what behaviors or outcomes to consider when making their ratings. In order for FOR to be effective, the rating scales should define performance levels as precisely as possible. Research shows that FOR can help to create this common observational "frame of reference" or schema and increase rater accuracy.[65]

Frame of reference training increases accuracy

Intentional bias

Raters may commit rating errors *intentionally* for political reasons or to provide certain outcomes to their employees or themselves.[66] For example, the most common intentional rating error in organizations is probably leniency. Managers may assign higher ratings than an employee deserves to avoid a confrontation with the employee, to protect an employee suffering from personal problems, to acquire more recognition for the department or themselves, to comply with organizational norms, to promote an employee out of a unit, or to be able to reward the employee with a bonus or promotion. Although less common, managers may also intentionally assign more severe ratings than an employee deserves to motivate him or her to work harder, to teach the employee a lesson, or to build a case for firing the employee. This is not considered a common or chronic problem for organizations although it would not be pleasing for the recipient of such deflated (and apparently severe) ratings. There is evidence that the error of leniency can be reduced by training raters on how to provide negative feedback and by holding raters more accountable for their rating tendencies.[67]

Other attempts to control intentional rating errors and biases include hiding scoring keys such as through forced choice, a forced distribution, forced ranking or other form of ratee comparison system, requiring cross-checks or reviews of ratings by other people, using multirater systems, training raters on how to provide negative evaluations, and reducing the rater's motivation to assign inaccurate ratings. **Unfortunately, none of these methods has proven to be reliably effective for controlling deliberate errors and biases.**[68]

There are a wide variety of PA training programs available for purchase, some on the Internet. One of the more effective programs for supervisors is "Legal and Effective Performance Appraisals," which takes the supervisors from PA preparation through the post-PA interview process. The highlights of this program, which should be covered in any comprehensive PA training program, are summarized in Figure 7-11. Remember also that training raters, and particularly the use of frame-of-reference training, is among the prescriptions that employers should follow to increase their chances of winning legal challenges related to performance-based decisions.

Most experts contend that the best ways to control for deliberate bias on the part of an individual rater are to hold raters more accountable for their ratings and to use more than one rater.[69] In general, the mean rating compiled from ratings across all (or a sample) of qualified raters will result in less bias and more validity for the performance appraisal system. A "qualified" rater can be defined as any internal or external customer who is the recipient of the performers' products or services.[70] The next section describes multirater (or 360-degree) appraisal systems.

Hold raters accountable

Figure 7-11 Legal and Effective Performance Appraisals: A Training Program for Supervisors

SAFEGUARDS AGAINST BIAS

1. Clearly communicate performance standards
   - Avoid subjective judgment-trait language on forms and in feedback
   - Expectations should be clearly understood with measurement precision
   - Standards should be fair and equitable (think disparate treatment discrimination)
2. Knowledge of PA procedures
   - Review evaluations with supervisor(s) before meeting
   - Allow employees to read, review, and sign off on performance appraisals
   - Have an appeal process—allow procedure for re-evaluation.
3. Linking PAs to job description detail is a key to effective PA
   - Good, up-to-date job descriptions facilitate clear understanding of tasks, responsibilities; they further long-term strategic goals of organization
   - Write accurate, up-to-date job descriptions
   - Set clear standards for rating job performance
   - Get job occupant input on job description and standards and sign-off

STEPS IN THE PA PROCESS

1. Preparation—How does the supervisor prepare?
   1. Gather documentation
   2. Review performance log/diary, incident reports, important information
   3. Review attendance records
   4. Review goals/expectations
   5. Review PA form
2. Encourage self-evaluation (but remember anchoring!)
   - Review self-evaluation after your initial appraisal
3. Set convenient time and place for uninterrupted meeting
4. Rate performance—typical performance level
   - Use behavioral/results/outcomes as criteria
   - Beware of rating errors (e.g., halo/horns; recency; leniency error; actor/observer bias)
5. Evaluate yourself as a manager and facilitator of performance
   - Consider constraints beyond performer's control

CONDUCTING THE PA INTERVIEW

1. Put employee at ease
   - Intention—collaborative, horizontal communication
   - Avoid negativity as much as possible
   - Attention to: Job-related, objective behaviors and countable results/outcomes/work products
     - Not personality traits or the person's characteristics
     - Remember: the focus is on *performance* (not traits)
2. Reach agreement on solutions and methods for improvement
   - Feedback should be behavioral/outcome/results-based (e.g. Don't say someone is "unreliable"; comment on the specific behavior or outcomes with as much precision as possible (define "unreliable")
   - Key to effective feedback is presenting the information in a way that prevents or lowers the probability of emotional reaction
   - Concentrate on observed behavior/results/the record of performance outcomes
3. Set goals for next PA period
   - Employees should have a say in setting their goals
   - GOALS SHOULD BE:
   1. Realistic (attainable)
   2. Motivating
   3. Contribute to productivity and compatible with strategic goals

POST-PA MEETING

- Do final evaluation after considering new information and self-evaluation
- Employee should sign and date form; provide opportunity to comment

EFFECTIVE APPRAISALS ARE AN ONGOING PROCESS—EMPLOYERS AND EMPLOYEES NEED REGULAR COMMUNICATION AND FEEDBACK TO DEVELOP TRUST AND SHARED COMMITMENT

Source: "Legal and Effective Performance Appraisals." Available from Coastal Technologies (http://econ.coastal.com).

Defining the Rater

Ratings can be provided by ratees, supervisors, peers, clients or customers, or high-level managers. While most companies still give the supervisor the sole responsibility for the employee's appraisal, formal multirater systems are becoming quite popular.[71] A growing number of companies use formal self-assessments.[72] Upward appraisals (ratings by subordinates) are also on the increase as part of a manager or supervisory PM&A system. Peer ratings have proven to be particularly valuable sources of information about performance and for judgments of a person's potentiality for future performance.[73]

We also know that the traditional single rater "top-down" approaches to PA are not very effective. Research has determined that so-called idiosyncratic variance in ratings (i.e., variability due to the particular raters who did the ratings) is often more related to the variability in performance ratings than the actual performance (this is not a good thing).[74] In other words, the particular rater who does the ratings unfortunately has too much to do with the ultimate rating that a particular level of performance is rated. We know, for example, that a rater's personality is related to particular rating tendencies such as leniency.[75] This problem obviously makes fair comparisons of rating data across raters (or supervisors) very difficult.

With increasing frequency, organizations are concluding that multiple rater types are beneficial for use in their appraisal systems.[76] Ratings collected from several raters, also known as **360-degree appraisal** systems, are thought to be more accurate and have fewer biases, are perceived to be more fair, and are less often the targets of lawsuits.[77] The use of 360-degree appraisal systems has also been identified as a **high-performance work practice** and thus linked to superior corporate financial performance. There are numerous web-based systems of 360-degree appraisal, some based on competency-based models of HR strategy.[78]

Multi-rater systems are a high-performance work practice

The probable reason that multirater appraisal is successful is that many of the rater types used (e.g., customers, peers) have direct and unique knowledge of at least some aspects of the ratee's job performance and can provide reliable and valid performance information on some job activities. In fact, the use of raters who represent all critical internal and external customers contributes to the accuracy and relevance of the appraisal system.[79]

Many organizations use self-, subordinate, peer, and superior ratings as a comprehensive appraisal prior to a training program. The Center for Creative Leadership in Greensboro, North Carolina, requires all participants in its 1-week assessment center program to first submit evaluations from superiors, peers, and subordinates. The data are tabulated by the center, and the feedback is reported to participants on the first day of the assessment center program. Participants consider this feedback to be among the most valuable they receive.

Many companies now use external customer data as an important source of information about employee and unit performance and for reward systems. The Marriott Corporation places considerable weight on its customer survey data in the evaluation of each hotel as well as work units within the hotels. Burger King, McDonald's, Domino's Pizza, and Taco Bell are among the companies that hire professional "customers" or "mystery shoppers" to visit specific installations to provide detailed appraisals of several performance functions.[80] Critical Thinking Application 7-B focuses on this approach to appraisal. Technology now provides better data on all kinds of customer factors that can be used for HR functions, especially pay-for-performance.[81]

Mystery shoppers

Figure 7-12 presents a summary of recommendations for implementing a multirater/360-degree appraisal system. There is no doubt that multirater PM&A increases the amount of information about a performer and provides very different perspectives on performance (the average correlations between subordinate and self-ratings and subordinate and supervisory ratings are only .14 and .22, respectively).[82] But in the context of the strategic objectives of the organization, the supervisor is probably the best source of information for making appraisals with this critical focus in mind. In addition, the immediate supervisor is probably the person most responsible for linking PA data to critical personnel decisions such as pay raises and terminations. However, there are many jobs in which the supervisor has few (if any) opportunities to observe performance. Gathering data from the critical internal and external customers is ideal for these situations.

Figure 7-12 Recommendations for Implementing a 360-Degree Appraisal System

INSTRUMENT ISSUES

- Items should be directly linked to effectiveness on the job.
- Items should focus on specific, observable behaviors and/or outcomes (not traits, competencies).
- Items should be worded in positive terms, rather than negative terms. Raters, particularly employees, may be less likely to respond honestly to negative items about their boss.
- Raters should be asked only about issues for which they have firsthand knowledge (i.e., ask subordinates about whether the boss delegates work to them; don't ask peers since they may not know).

ADMINISTRATION ISSUES

- Select raters carefully by using a representative sample of people most critical to the ratee (and the work unit) and who have had the greatest opportunity to observe his or her performance.
- Use an adequate number of raters to ensure adequate sampling and to protect the confidentiality of respondents (at least three per source; except supervisor). An alternative strategy is to solicit ratings from all possible qualified raters.
- Instruct respondents on how the data will be used and ensure confidentiality.
- To maintain confidentiality, raters should not indicate their names or other identifying characteristics and surveys should be returned in a manner so as to maintain confidentiality.
- Alert and train raters regarding rating errors (e.g., halo, leniency, severity, attributional bias).

FEEDBACK REPORT

- Separate the results from the various sources. The ratee should see the average, aggregated results from peers, subordinates, higher-level managers, customers, or other sources that may be used.
- Show the ratee's self-ratings as compared to ratings by others. This enables the ratee to see how his or her self-perceptions are similar to or different from others' perceptions.
- Compare the ratee's ratings with other norm groups. For example, a manager's ratings can be compared to other managers (as a group) in the firm.
- Provide feedback on items as well as scales so ratees can see how to improve.

FEEDBACK SESSION

- Use a trained facilitator to provide feedback to ratees.
- Involve the ratee in interpreting his or her own results.
- Provide an overview of the individual's strengths and areas for improvement.
- Provide feedback on recommendations and help him or her to develop an action plan.

FOLLOW-UP ACTIVITIES

- Provide opportunities for skill training in how to improve his or her behaviors.
- Provide support and coaching to help him or her apply what has been learned.
- Over time, evaluate the degree to which the ratee has changed behaviors.

Source: Modified from G. Yukl and R. Lepsinger, "360 Feedback," *Training,* December 1995, pp. 45–48, 50.

Even if an organization doesn't use a formal multirater system, many supervisors should (and do) use indirect information and may alter their ratings based on information that they did not personally observe. In general, data from multiple sources are recommended because they provide a more comprehensive "picture" of an individual's performance and contribution.[83]

Defining the "Ratee"

Just like particular rater characteristics can have an impact on ratings, so too can the individual characteristics of those who are rated have an impact on ratings in addition to the actual performance levels of these individuals. The good news from this abundant literature is, despite the influence of the particular rater and his or her proclivities, the actual performance and a person's ability level tend to have the highest correlations with resultant performance ratings. More precisely defined performance measures will help control or reduce the extent to which irrelevant personal ratee characteristics, such as race, gender, or age, enter into the rating process. Research also indicates that multirater systems can reduce these potential sources of bias as well.[84]

Rating the Unit on Performance

Many people assume that appraisals always focus on an *individual* level of performance. There are alternatives to using the individual as the ratee that are becoming more common

in organizations as more firms (e.g., General Foods Corporation, Rohm & Haas, General Motors, Westinghouse) shift to using more **self-managed teams and other team-based organizational structures.** Thus, PM&A systems should assess overall team performance along with (if possible) individual team members' contributions to team performance. Thus, the object of appraisal can be defined at the individual, work group, division, or organization wide level. It is also possible to define the object (or ratee) at multiple levels. For example, for some performance dimensions, it may be desirable to appraise performance at the work group level for merit pay purposes and additionally at the individual level to identify particular developmental needs for team members. Burger King, for example, awards cash bonuses to branch stores based on a customer-based evaluation process while maintaining an individual appraisal system within each store. Delta Airlines assesses customer service at the unit level only, while other job activities are assessed at the individual employee level.

Team PAs for high work group cohesiveness

Two conditions that make it desirable to assess performance at a higher aggregation level than the individual level are high work group cohesiveness and difficulty in measuring individual contributions. **High work group cohesiveness** refers to the shared feeling among work group members that they form a team. Such an orientation promotes high degrees of cooperation among group members for highly interdependent tasks. Appraisals focused on individual performance may undermine the cooperative orientation needed to maintain this cohesiveness and tend to promote individualistic or even disruptive competitive environments. In some cases, workers are so **interdependent** (their individual performance outcomes cannot be clearly determined) that there is no choice but to focus their appraisals on the performance of their work group only.

These conditions do not rule out the possible measurement of individual performance in the team context. If individual performance is not measured in teams, the possibility of "social loafing" is more likely where team members tend to make less of an effort to achieve a goal when they work in a group versus when they work alone. To make matters worse, when other very capable team members determine that there are "free riders" (these are the "loafers"), they may withdraw their efforts toward team performance.[85]

Process tracing software

Technology now allows for the collection of more objective (and more valid) data on the levels of individual contributions to teams or projects. **Process tracing software** is now available and used by some companies to provide data on the interactions and contributions of individual team members. For example, Microsoft uses data from its software to identify programming "sparkplugs" (those who originate an idea), the "super-connectors" (those who build on an idea), and the "bottlenecks" (those who hold things up). It then uses these results to reward contributions and to plan future assignments. IBM uses similar software to identify employees who will be "fast-tracked" into other projects and other leadership roles based on their contributions to group projects.[86]

So our recommendation is to make a concerted effort to assess individual contributions to team performance. This can usually be achieved using peer assessment since peers are often in the best position to assess individual team members' contributions. Above all, it should be understood that not all work "teams" are the same, so a set of PM&A prescriptions for all "teams" will not work. An excellent summary of team-based PA concluded that **"effective performance appraisal is a matter of fit between characteristics of the team and the target of assessment, as well as the rating type, source, and purpose."**[87] Performance-appraisal systems that result from careful consideration of these contingencies have the greatest probability of being effective; that is, of eliciting employee behavior that contributes to an organization's goals.

Many companies rely on aggregated data to assess unit performance. One of the most popular and successful approaches is discussed next.

The Productivity Measurement and Enhancement System (ProMES)

One excellent approach to measuring aggregated performance is "The Productivity Measurement and Enhancement System" or **(ProMES).**[88] Similar to management by objectives but usually for aggregated, unit performance, the purpose of ProMES is to measure performance with the purpose of improving unit productivity and overall performance. The performance measurement system is developed by employees (with management approval),

and the feedback on the performance measures is then used to help the work unit improve. ProMES is designed to give workers the precise performance information they need to perform more effectively and to give them a sense of ownership and empowerment.

ProMES is designed to increase performance and productivity by improving motivation. While it can also be used as a management information system, the most important function is to provide feedback about productivity in order to help workers perform more effectively.

ProMES has an excellent track record

The approach has an excellent track record and has proven successful in a variety of settings. **Research with ProMES indicates that it is a highly effective method for improving performance while also improving job satisfaction and reducing job stress.**[89]

Figure 7-13 presents a summary of the major steps to follow in ProMES. The first step is to form a design team made up of employees from the work unit that will ultimately use the system. This design team, made up of from five to eight people, should include supervisors from the unit and also a ProMES facilitator. The design team must first come up with one set of objectives plus quantitative indicators to be used for feedback on these objectives. The objectives are derived from a study of the specific tasks that this unit must accomplish for the organization. For example, the objectives could be rather general statements such as "effectively dealing with production priorities" or "optimizing customer satisfaction" or "providing a safe working environment." Next, the quantifiable measures of performance or "indicators" are written. These indicators need to clearly stipulate how well all objectives are being met. To identify these indicators, the design team is asked to think of measures that show how well objectives are being met. There is at least one "indicator" for each objective. Examples of indicators might be a percentage of errors made, an average time between failures of repaired items, or a percentage of satisfied customers. There are usually from four to six objectives and from 8 to 12 performance indicators. Once the

Figure 7-13
THE STEPS OF ProMES

STEP 1. FORM THE DESIGN TEAM
- The people who will be primarily responsible for developing the measurement and feedback system (includes supervisor and facilitator)

STEP 2. IDENTIFY THE OBJECTIVES
- Group discussion leading to consensus

STEP 3. IDENTIFY INDICATORS
- Quantitative indicators developed for each objectives
- Indicators must be largely under the control of those being measured

STEP 4. DEFINE CONTINGENCIES
- Operationalize the product-to-evaluation contingencies
- Derive utility functions relating changes in the amount of the indicator (the product) to variation in unit effectiveness
- Defines how much of an indicator is how good for the organization
- Management reviews and approves contingencies

STEP 5. DESIGN THE FEEDBACK SYSTEM
- Regular (often monthly) computerized reports go to unit personnel
- Effectiveness score for each indicator value is provided
- Overall effectiveness score provided (+ historical data)
- Identifies priorities for improvement

STEP 6. PROVIDE CONTINUOUS FEEDBACK AND RESPOND
- Focus on individual indicators
- Discuss causes of improvements or decreases

STEP 7. MONITOR THE SYSTEM OVER TIME
- Make adjustments in ProMES measurement systems
- Particularly important if indicators are new

Source: Pritchard, R. D., Harrell, M. M., DiazGranados, D., & Guzman, M. J. (2008). The productivity measurement and enhancement system: A meta-analysis. *Journal of Applied Psychology, 93*, 540–567.

consensus-driven objectives and indicators emerge from the work of the design team, they are reviewed and, perhaps after a few iterations, ultimately approved by management. The goal of management is to ensure that the objectives and the indicators are aligned with broader goals or objectives of both the unit and the organization. Figure 7-14 describes objectives and indicators derived from the design team for a circuit board production unit.

The next step is for the design team to consider and define "contingencies." These "contingencies" are a form of graphic utility function that relates an amount or measure of an indicator to a value for the organization. A hospital might use a percentage of bed capacity in the intensive care unit as an indicator. One axis of the utility function would show an indicator range level, and the other axis would represent *"effectiveness levels"* or the

Figure 7-14 Examples of Objectives and Indicators

ORGANIZATIONAL CONSULTANTS

Setting: This unit worked with clients doing individual assessments of various types ranging from one-day assessment to multiple-day assessment centers.

Objective 1. Profitability

- Indicator 1. Cost Recovery. Average amount invoiced per assessment divided by cost for that assessment.
- Indicator 2. Billable Time. Percent monthly billable time on days when any assessment function is done.
- Indicator 3. Billing Cycle Time. Average number of days between billing trigger and invoice submission.

Objective 2. Quality of Service

- Indicator 4. Validity of Selection Assessments. Percentage of hits: people assessed predicted to be high performers who turn out to be high performers and those predicted to be marginal who are marginal. Index is based on a 6 - month follow up.
- Indicator 5. Cycle Time. Percentage of assessment reports going out that went out on time.
- Indicator 6. High Quality Experience of Participant. Percentage of participants giving "Satisfied" and "Very Satisfied" ratings at the time of assessment.
- Indicator 7. Customer Satisfaction. Percentage of "Satisfied" and "Very Satisfied" to customer satisfaction measure.
- Indicator 8. Consultant Qualifications. Percent licensable consultants who are licensed within two years of joining the firm.
- Indicator 9. Ethics/Judgment Training. Percent staff with a minimum of 4 hours ethics/judgment training in the last 12 months.

Objective 3. Business Growth

- Indicator 10. Assessment Revenue. Average revenues for the last three months from the assessment function.

Objective 4. Personnel Development and Satisfaction

- Indicator 13. Personnel Skill Development. Number of actual tasks the person had been trained on divided by the number of possible tasks that person could be trained on.
- Indicator 14. Personnel Satisfaction. Average number of "OK" and "Good" days per person per month based on data entered when each person entered his/her weekly time card.

PHOTOCOPIER REPAIR PERSONNEL

Setting: Technicians go out on service calls to repair customers' photocopiers.

Objective 1. Quality: Repair and maintain photocopiers as effectively as possible.

- Indicator 1. Mean copies made between service calls
- Indicator 2. Percentage repeat calls
- Indicator 3. Percentage of preventive maintenance procedures correctly followed

Objective 2. Cost: Repair and maintain photocopiers as efficiently as possible.

- Indicator 4. Parts cost per service call
- Indicator 5. Labor time per service call
- Indicator 6. Percentage of repeat service calls caused by a lack of spare parts

Objective 3. Administration: Keep accurate records of repair and maintenance

- Indicator 7. Percentage of required repair history information filled in correctly
- Indicator 8. Percentage of parts warranty claims correctly submitted.

Objective 4. Attendance: Spend the available work time on work related activities.

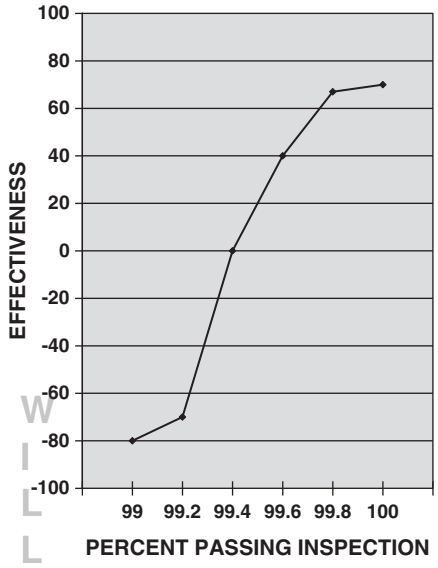
- Indicator 9. Percentage of labor contract hours actually spent on the job.

Objective 5. Ambassadorship: Behave as correctly as possible on the job.

- Indicator 10. Percentage of important social behaviors shown on the job as measured by customers' ratings.

Source: Pritchard, R. D., Weaver, S. J. & Ashwood, E. L. (2012). *Evidence-based productivity improvement: A practical guide to the Productivity Measurement and Enhancement System.* New York: Routledge, Taylor & Francis Group.

Figure 7-15
Function Table Contingency



Source: Pritchard, R. D., Weaver, S. J. & Ashwood, E. L. *(2012). Evidence-based productivity improvement: A practical guide to the Productivity Measurement and Enhancement System.* New York: Routledge, Taylor & Francis Group.

amount of contribution a certain indicator level is related to the organization's objectives. This utility function then defines how each level of the indicator is related to effectiveness. A contingency function must be generated for each indicator. Figure 7-15 presents a contingency table for the circuit board production unit indicator.

Identify relative importance of indicator

Part of this process involves identifying the realistic maximum and minimum levels for each indicator and reaching consensus on the minimum level of acceptable performance on each indicator (i.e., just meeting minimum expectations). This minimum level means that falling below this point would represent performing below minimum expectations on the indicator. The group also ranks and rates the effectiveness levels of the maximum and minimum indicator levels for each indicator. The result will be an effectiveness score for the maximum and minimum indicator levels for each contingency. This process identifies the relative importance of each indicator, the particular quantitative levels of performance for each indicator (indicators will have different ranges of performance), and the points where changes in indicator levels won't necessarily translate into the same amount of change in effectiveness. For example, in the intensive care unit, the process revealed that an increase in bed capacity above 75 percent was not very valuable. **The other advantage of this step is that it helps to identify priorities for improvement.** Thus, the gain in effectiveness can be measured if the unit improved on a particular indicator. For example, an improvement from 70 to 75 percent bed capacity means a gain in effectiveness of +60 points, whereas a gain from 75 to 80 percent would represent a +20 gain. This indicates that improving bed capacity would be a high priority below 75 percent but a much lower priority when above 75 percent. Also, since all contingencies for each indicator scale on the common "effectiveness" metric, an overall effectiveness score can be formed for the unit by summing the effectiveness scores for each indicator. **This overall effectiveness score then provides the index of overall productivity.** Of course, management also reviews all contingencies, the functions and definitions, and the minimum acceptable levels of performance.

After all contingencies are approved by management, the data collection and feedback system is then installed. Each unit member receives (via computer) a feedback report (usually monthly) that includes the list of the unit objectives, the indicators, the performance level on each indicator, and the effectiveness scores. A feedback meeting is then held in which the unit members and the supervisors review the report and try to identify steps to be taken to improve performance in particular areas.

A recent meta-analysis found that ProMES results in significant improvements in productivity, that the effects tend to last over time (in many cases, years), and that the

improvements occur in many different types of organizational settings that differ on the type of organization, the type of work performed, the types of workers, and the country where the ProMES intervention occurs.[90]

Research on ProMES is strong

While ProMES is typically done with work units, the approach is adaptable (and has been used) for individuals, and combinations of individual and group measures can also be used. ProMES can also be part of a "benchmarking" process where performance measures for similar work units can be compared across organizations. *Benchmarking* is one example of a process whereby a particular unit can evaluate its performance relative to some other comparable unit, either inside or outside the organization. We take up the issue of benchmarking next.

Benchmarking—Gauging internal practices to external standards

Benchmarking is the process of gauging the internal practices and activities within a firm to an external reference or standard. It is a continuous data-driven process of measuring one's own products, services, systems, and practices against the world's toughest competitors to identify areas for improvement. Although the approach could be used for individual performers, it is most often used to evaluate unit-level data.

Survey data is positive

A recent survey found that some form of benchmarking was used by a majority of respondent organizations but that so-called best practice benchmarking is used by less than half of benchmarking organizations.[91]

Ford Motor Company benchmarked its accounts payable function against Mazda Motor Corporation. Ford found that it had about five times as many employees as it needed. The automaker redesigned the system for tracking orders, deliveries, and invoices and thereby helped employees to perform the same tasks more efficiently. As a result, Ford was able to simplify the process, reduce the number of employees, and reduce errors. Goodyear Tire and Rubber changed its compensation practices by benchmarking what several Fortune 100 firms were doing in compensation. It developed a system to link employee performance to the firm's financial gains. AT&T examined the role of chief financial officers to redesign the job duties and functions of its CFO to be more in line with what world-class CFOs were doing.

Needs top management support

Studies on the effectiveness of benchmarking have found that it is critical to have top management support and commitment to the process, including the "benchmarked" companies. In addition, when it results in setting moderately difficult goals that employees believe are attainable, it seems to work. But when poorer-performing companies receive benchmarking data that their practices are significantly different from the "best practices," and their managers set radical, unrealistically high goals, employees have difficulty embracing the changes and may resist them. As a result, performance actually may decline.

Recent survey research indicates that the perceived effectiveness of benchmarking compares favorably with the effectiveness of most intervention strategies (but less so than business process reengineering, quality management systems, and customer satisfaction data). A majority of respondents indicated that they intend to continue using benchmarking in the future.[92] Perhaps setting more realistic goals and gradually increasing the difficulty of the goals will encourage employees. This process is known as **shaping,** which is a behavioral change technique that promotes gradual improvement from a known, initial behavior to a desired goal, or, in this case, the benchmark. For example, if an organization wants to meet the best practice of having 1 percent defects in its industry, and its initial performance is at 20 percent defects, the company may need to first use 15 percent defects as a goal. Once workers master that goal and are rewarded, then the company can change the goal to 10 percent defects and so on. In this way, the company is continually moving toward the benchmark goal and employees are less resistant than if they were initially assigned the goal of 1 percent defects, which they may have felt was unattainable. To use shaping effectively in benchmarking practices, the following tips are offered.

1. Identify what is to be benchmarked (a process, product, service, etc.).
2. Identify comparable companies.
3. Collect data to precisely define the target goal (benchmark).
4. Collect data to determine the organization's current performance level against the benchmark.
5. Reduce the target to discrete, measurable, smaller steps or goals.

6. Train, as needed, any employees so that they can meet the smaller goals (subgoals).
7. Periodically provide feedback and use appropriate, valued reinforcers for meeting the subgoals.
8. Increase the subgoals so that they are getting closer to the target goal.
9. Recalibrate benchmarks periodically.

The recalibration is important so that the organization continually monitors the benchmark or target goal because it may change. Successes by companies may lead to new standards.

Benchmarking should be considered one form of performance measurement that provides a basis of comparison to competitors and other outside sources. While this is a useful approach to measurement, the importance attached to any measurement should derive from the extent to which the measurement is related to the strategic goals of the organization.

Administrative Characteristics

Figure 7-16 presents a summary of the many issues to consider regarding the administrative characteristics of a PM&A system. Among the most important characteristics are the extent to which computers are used to make and maintain ratings and the methods of delivering feedback.

Almost all PM&A systems discussed previously are now computer adaptive and some require it (e.g., CARS, PDA). There are now several online systems of 360-degree appraisal that are used by many of the most successful companies of the world. The reader should consult one of the following websites to sample online 360-degree systems: PersonnelDecisions.com, 360-degreefeedback.com, performaworks.com, acumen.com, cwginc.com, or fullcirclefeedback.com.

Performance Monitoring

One administrative issue is the possible automation of performance measurement. Can we eliminate the raters altogether? The practice of monitoring employees while they perform their jobs through the use of surveillance cameras, telephone monitoring, or computer monitoring is growing in popularity. Remember the discussion of the JetBlue at-home (in Utah) reservationists? Do you think they can slip away from their CSR duties and do a little Facebooking? Not a chance. JetBlue has an elaborate performance monitoring system that records everything important about each reservationist's performance during on-duty time. An automated system even tells the employee when to take breaks. More companies are turning to some form of monitoring regarding workers' online behavior. They probably should. The reported rates of on-the-clock, online cruising are rather alarming. Most companies maintain that such performance monitoring is an acceptable and ethical means for gathering information about performance and other aspects of work. Information from

Figure 7-16
Major Administrative Issues to Consider in Performance Management

1. Frequency and timing of formal appraisals
   - Number of times per year (e.g., one per year, every 6 months, quarterly?)
   - Time period (e.g., anniversary of hire, after project completion)
2. Rating/data collection medium
   - Computerized data collection/data tabulation/integration into database
   - Hard copy for personnel file and sign off?
   - Use of technology for performance data collection and monitoring
   - Computer programs that can monitor rater rating tendencies
3. Training programs
   - For raters (supervisors), ratees, administrators
   - Scheduling/assessment/follow-up
   - Frame of Reference (FOR)/self-efficacy training
4. Method of feedback
   - Feedback via computer versus scheduled sessions
   - Feedback based on comparisons to other employees/companies
   - Formal feedback sessions with supervisors, team, consultants, coaches

electronic monitoring should be incorporated into the full performance management system.

Employees in general don't like most electronic monitoring even when the monitoring can result in positive outcomes for the ratees. Offering those who are to be monitored input into the monitoring process reduced invasion of privacy concerns while team leaders are more likely to monitor performance in secret when there is a low level of trust in a work group. In addition, team leaders tend to increase their level of electronic monitoring over time.[93]

Methods of Delivering Performance Feedback

Supervisors or managers should communicate appraisal results to ratees through a formal feedback "PA" meeting held between the supervisor and the employee(s).[94] Feedback serves an important role both for motivational and informational purposes and for improved rater–ratee communications.[95] Recall the ProMES process described earlier. For example, **supportive feedback can lead to greater motivation,** and feedback discussions about pay and advancement can lead to greater employee satisfaction with the process.

Provide specific and timely feedback

Detailed and specific feedback (e.g., "this book uses too many rambling sentences and big words") is recommended instead of general feedback ("I hate the writing") since **more precision is more likely to improve performance.**

A meta-analysis found that four feedback characteristics were related to performance improvements after feedback. The largest effects from feedback occurred when the working tasks were more familiar to the ratee, there were performance cues that supported learning and improvement, the feedback provided information on discrepancies between performance and a precisely defined performance standard, and the feedback did not threaten the ratee psychologically.[96]

The biggest hazard for the rater in providing performance feedback may be ratee reactions to the feedback. Generally, ratees believe that they perform at higher levels than do observers of that performance.[97] This is especially true at the lower performance levels where there is more room for disagreement and a greater motive on the part of ratees to engage in ego-defensive behavior. Let's not forget about the **actor-observer bias** factor either. It is no wonder that raters are often hesitant about confronting poor performers with negative appraisal feedback and may be lenient when they do. Although pressure on managers to give accurate feedback and to effect change may override a reluctance to give negative feedback, the pressure doesn't make the experience any more pleasant. In addition, feedback to inform poor performers of performance deficiencies and to encourage improvement doesn't always translate into higher performance.[98]

Accurate feedback doesn't always help

Recommendations for feedback sessions

To create a supportive atmosphere for the feedback meeting between the employee and supervisor, several recommendations exist. Raters should avoid being disturbed and should take sufficient time in the meeting. They should keep notes on effective and ineffective behavior as it occurs so that they will have some notes to refer to when conducting the feedback session (review the legal prescriptions presented earlier). Raters should be informal and relaxed and allow the employee the opportunity to share his or her insights. Topics that should be addressed include praise for special assignments, the employee's own assessment of his or her performance, the supervisor's response to the employee's assessment, action plans to improve the subordinate's performance, perceived constraints on performance that require subordinate or supervisory attention, and employee career aspirations, ambitions, and developmental goals. In sum, raters should provide feedback that is clear, specific, descriptive, job related, constructive, frequent, and timely. Recipients of the feedback are more likely to perceive the information as accurate and agree to attend to shortcomings when the feedback is derived from multiple rater systems that involve internal and external customers.[99]

Strength-based PA

Another promising avenue for improving the effectiveness of performance appraisal feedback is called **"Strength-Based Performance Appraisal."** Combined with goal-setting, this approach puts the focus on existing worker strengths while constraining negative feedback by concentrating on prevention-focused behaviors. Preliminary evidence on the approach is positive.[100]

SUMMARY

Despite popular but unconvincing arguments to the contrary, performance appraisals remain an important tool for organizations to manage and improve the performance of employees and work units, to make more valid staffing decisions, and to enhance the overall effectiveness of an organization's services and products. The design, development, and implementation of appraisal systems are not endeavors that can be effectively handled by following the latest fad or even by simply copying other organizations' systems. Instead, a new PM&A system must be considered a major organizational change effort that should be pursued in the context of improving the organization's competitive advantage. This means that, like any such change effort, there will be vested interests in preserving the status quo that will resist change, no matter how beneficial it may be for the organization. These sources of resistance to the change have to be identified and managed to build incentives for using a new appraisal system. We are impressed with the ProMES method, its apparent utility in many diverse settings, and the research indicating its effectiveness. If the main purpose of the PM&A system is to improve performance, this approach should be considered.

Once a well-designed system has been implemented, the work is still not done. A PM&A system has to be maintained by monitoring its operation through periodic evaluations. Only by keeping a PM&A system finely tuned will managers have a rational basis for making sound personnel decisions to achieve the kinds of gains in productivity that are so critically needed in today's times. PM&A should be an integral part of the strategic HR system. Data from this system should be a critical component for all sorts of internal staffing decisions (promotions, retentions, terminations, pay).

Among the personnel decisions, some of the most important concern the organization's compensation system. The prescriptions presented in Figure 7-1, the findings discussed in Figure 7-2, and the training recommendations we have presented should be helpful guidelines for improving most PM&A systems. Effective PM&A also must be carefully integrated with other human resource domains, particularly compensation systems with a pay-for-performance component. Accurate appraisals also are critical for determining training needs, one of the subjects of the next chapter.

Discussion Questions

1. Why has performance appraisal taken on increased significance in recent years?
2. As the workforce becomes more diverse, why does performance appraisal become a more difficult process?
3. Ford was accused of age discrimination based on the use of its forced-distribution rating system. What evidence would you investigate to test this allegation?
4. Many managers describe performance appraisal as the responsibility that they like the least. Why is this? What could be done to improve the situation?
5. Describe several advantages and disadvantages to using rating instruments that are based on comparisons among ratees' performance, comparisons among anchors, and comparisons to anchors.
6. What steps would you take if your performance appraisal system resulted in disparate or adverse impact?
7. Under what circumstances would you use customer or client evaluation as one basis for appraising employees?
8. Why are so many companies using 360-degree feedback systems? What are the benefits of such systems?
9. Why should managers provide ongoing and frequent feedback to employees about their performance?
10. As an employee, how would you react to a forced-distribution rating system?

WILLIS, KASSANDRA 2161TS