16

TRAINING AND DEVELOPMENT

Implementation and the Measurement of Outcomes

LEARNING GOALS

By the end of this chapter, you will be able to do the following:

16.1 Classify training methods as presentation, hands-on, group building, or technology based

16.2 Identify key principles of instructional design to encourage active learner participation

16.3 List the essential elements for measuring training outcomes

16.4 Explain the advantages and disadvantages of ROI and utility analysis as methods for assessing the financial outcomes of learning and development activities

16.5 Identify potential contaminants or threats to valid interpretations of findings from field research

16.6 Distinguish experimental from quasi-experimental designs

16.7 Describe the limitations of experimental and quasi-experimental designs

16.8 In assessing the outcomes of training efforts, distinguish statistically significant effects from practically significant effects

Once we define what trainees should learn and what the substantive content of training and development should be, the critical question then becomes "How should we teach the content and who should teach it?"

The literature on training and development techniques is massive. Although many choices exist, evidence indicates that, among U.S. companies that conduct training, few make any systematic effort to assess their training needs before choosing training methods (Arthur, Bennett, Edens, & Bell, 2003; Saari, Johnson, McLaughlin, & Zimmerle, 1988). In a recent survey, for example, only 15% of learning and development professionals reported that they

lack data and insights to understand which solutions are effective (LinkedIn Learning, 2017). This implies that firms view hardware, software, and techniques as more important than outcomes. They mistakenly view the identification of what trainees should learn as secondary to the choice of technique.

New training methods appear every year. Some of them are deeply rooted in theoretical models of learning and behavior change (e.g., behavior modeling, team-coordination training), others seem to be the result of trial and error, and still others (e.g., interactive multimedia, computer-based business games) seem to be more the result of technological than of theoretical developments. In 2016, 41% of learning hours used at the average organization were delivered by technology-based methods, which is nearly 10 percentage points higher than in 2008 and more than 15 percentage points higher than in 2003. Technology-based learning can be delivered through online or other live remote classrooms, self-paced online or nonnetworked computer-based methods, mobile devices (such as smartphones and tablets), or noncomputer technology (such as DVDs) (Association for Talent Development, 2016).

We make no attempt to review specific training methods that are or have been in use. Other sources are available for this purpose (Goldstein & Ford, 2001; Noe, 2017; Wexley & Latham, 2002). We only highlight some of the more popular techniques, with special attention to technology-based training, and then present a set of criteria for judging the adequacy of training methods.

CATEGORIES OF TRAINING AND DEVELOPMENT METHODS

Following Noe (2017), training and development methods may be classified into four categories: presentation, hands-on, group building, and technology based.

Presentation Methods

With presentation methods, an audience typically receives one-way communication from the trainer in one of two formats:

- Lectures
- Videos, usually used in conjunction with lectures to show trainees real-life experiences and examples

Hands-On Methods

Hands-on methods include on-the-job training, self-directed learning, apprenticeships, and simulations:

- *On-the-job training*: New or inexperienced employees learn in the work setting and during work hours by observing peers or managers performing a job and then trying to imitate their behavior (Tyler, 2008). Examples include onboarding, job rotation, understudy assignments (also known as "shadowing," in which an understudy relieves a senior executive of selected responsibilities, thereby allowing him or her to learn certain aspects of the executive's job; see Dragoni, Park, Soltis, & Forte-Trammell,

2014), and executive coaching. Executive coaching is an individualized process of executive development in which a skilled expert (coach) works with an individual who is in a leadership or managerial role in an organization, to help the individual to become more effective in his or her organizational roles(s) and contexts (Vandaveer, 2017; see also Hollenbeck, 2002; Peterson, 2011; Underhill, McAnally, & Koriath, 2008).

- *Self-directed learning:* Trainees take responsibility for all aspects of learning, including when it is conducted and who will be involved. Trainers may serve as facilitators, but trainees master predetermined content at their own pace.
- *Apprenticeship:* This method constitutes a work-study training regimen that includes both on-the-job and classroom training. It typically lasts an average of four years.
- *Simulations:* These training methods represent real-life situations, with trainees' decisions resulting in outcomes that reflect what would happen if they were on the job. Simulations may assume a number of forms, including the following:
  - In the *case method,* representative organizational situations are presented in text form, usually to groups of trainees who subsequently identify problems and offer solutions. Individuals learn from each other and receive feedback on their own performances.
  - The *incident method* is similar to the case method, except that trainees receive only a sketchy outline of a particular incident. They have to question the trainer, and, when they think they have enough information, they attempt a solution. At the end of the session, the trainer reveals all the information he or she has, and trainees compare their solutions to the one based on complete information.
  - *Role playing* includes multiple role-playing, in which a large group breaks down into smaller groups and role plays the same problem within each group without a trainer. All players then reassemble and discuss with the trainer what happened in their groups.
  - *Experiential exercises* are simulations of experiences relevant to organizational psychology. This is a hybrid technique that may incorporate elements of the case method, multiple role-playing, and team-coordination training. Trainees examine their responses first as individuals, then with the members of their own groups or teams, and finally with the larger group and with the trainer.
  - The *task model* has trainees construct a complex, but easily built physical object, and a group of trainees must then duplicate it, given the proper materials. Trainees use alternative communication arrangements, and only certain trainees may view the object. Trainees discuss communication problems as they arise, and they reach solutions through group discussion.
  - The *in-basket technique* (see Chapter 13).
  - *Business games* (see Chapter 13).
  - *Assessment centers* (see Chapter 13).
  - *Behavior or competency modeling* (see Chapter 15).

Group-Building Methods

Group-building training methods are designed to improve group or team effectiveness. They include the following types of training:

- *Adventure learning:* This experiential learning method focuses on the development of teamwork and leadership skills through structured activities. These may include wilderness training, outdoor training, improvisational activities, drum circles, even cooking classes (Noe, 2017). Their purpose is to develop skills related to group effectiveness, such as self-awareness, problem solving, conflict management, and risk taking (Greenfield, 2015).

- *Team training:* This method is designed to improve effectiveness within the many types of teams in organizations (production teams, service teams, project teams, management teams, and committees; see Chapter 15). It focuses on improving knowledge (mental models that allow trainees to function well in new situations); attitudes (beliefs about a team's task and feelings toward team members); and behavior (actions that allow team members to communicate, coordinate, adapt, and complete complex tasks to accomplish their objective) (Salas, Burke, & Cannon-Bowers, 2002).

- *Action learning:* In this method, teams work on actual business problems, commit to an action plan, and are responsible for carrying out the plan (Malone, 2013; Pedler & Abbott, 2013). It typically involves 6–30 employees and may include customers or vendors as well as cross-functional representation. Teams are asked to develop novel ideas and solutions in a short period of time (e.g., two weeks to a month), and they are required to present them to top-level executives.

- *Organization development:* This method involves systematic, long-range programs of organizational improvement through action research, which includes (a) preliminary diagnosis, (b) data gathering from the client group, (c) data feedback to the client group, (d) data exploration by the client group, (e) action planning, and (f) action; the cycle then begins again. Although action research may assume many forms (Austin & Bartunek, 2003), one of the most popular is *survey feedback* (Church, Waclawski, & Kraut, 2001; Levinson, 2014; Wiley, 2010). The process begins with a comprehensive assessment of the way the organization is currently functioning—typically via the administration of anonymous questionnaires to all employees. Researchers tabulate responses at the level of individual work groups and for the organization as a whole. Each manager receives a summary of this information, based on the responses of his or her immediate subordinates. Then a change agent (i.e., a person skilled in the methods of applied behavioral science) meets privately with the manager recipient to maximize his or her understanding of the survey results. Following this, the change agent attends a meeting (face to face or virtual) of the manager and subordinates, the purpose of which is to examine the survey findings and to discuss implications for corrective action. The role of the change agent is to help group members to better understand the survey results, to set goals, and to formulate action plans for the change effort.

Technology-Based Training

Instructor-led, face-to-face, classroom training still comprises 49% of available hours of training (down from 64% in 2008), and if one considers all instructor-led delivery methods (classroom, online, remote), that figure rises to 65% of all learning hours available

(Association for Talent Development, 2016). The use of technology-delivered training is expected to increase dramatically, however, in the coming years as technology improves, its cost decreases, the demand increases for customized training, and organizations realize the potential cost savings from training delivered via tablets, smartphones, and social media. Currently, seven out of 10 organizations are incorporating video-based online training into their learning cultures, and 67% of people are learning on mobile devices (LinkedIn Learning, 2017).

Technology-based training creates a *dynamic learning environment*, it facilitates *collaboration*, and it enables *customization* (in which programs can be adapted based on learner characteristics) and *learner control*. That is, learners have the option of self-pacing exercises, exploring links to other material, chatting with other trainees and experts, and choosing when and where to access the training (Noe, 2017).

There are at least 15 forms of technology-based training (Noe, 2017):

- E-learning, online learning, computer-based training, and Web-based training
- Webcasts or webinars—live, Web-based delivery in dispersed locations
- Podcasts—Web-based delivery of audio- and video-based files
- Mobile learning—through handheld devices such as tablets or smartphones
- Blended learning—hybrid systems that combine classroom and online learning
- Wikis—websites that allow many users to create, edit, and update content and to share knowledge
- Distance learning—delivered to multiple locations online through webcasts or virtual classrooms, often supported by chat, e-mail, and online discussions
- Social media—online or mobile technology that allows the creation and exchange of user-generated content; includes wikis, blogs, networks (e.g., Facebook, LinkedIn), micro-sharing sites (e.g., Twitter), and shared media (e.g., YouTube)
- Shared workspaces, such as Google Docs, hosted on a Web server, where people can share information and documents
- RSS (real simple syndication) feeds—updated content sent to subscribers automatically instead of by e-mail
- Blogs—Web pages where authors post entries and readers can comment
- Micro-blogs or micro-sharing (e.g., Twitter)—software tools that enable communications in short bursts of texts, links, and multimedia
- Chat rooms and discussion boards—electronic message boards through which learners can communicate at the same or different times (a facilitator or instructor may moderate the conversations)
- Massive, open, online courses (MOOCs)—designed to enroll large numbers of learners (massive); free and accessible to anyone with an Internet connection (open and online); using videos of lectures, interactive coursework, including discussion groups and wikis (online); with specific start and completion dates, quizzes, assessments, and exams (courses)
- Adaptive training—customized content presented to learners based on their needs

Is technology-based training more effective than instructor-led training? Two meta-analyses have found no significant differences in the formats, especially when both are used to teach the same type of knowledge, declarative or procedural (Sitzmann, Kraiger, Stewart, & Wisher, 2006; Zhao, Lei, Lai, & Tan, 2005). Perhaps more important questions are these: How does one determine the optimal mix of formats for a program (e.g., blended learning), and does the sequencing of technology-based and in-person instruction within a program make a difference (Bell, Tannenbaum, Ford, Noe, & Kraiger, 2017)? Does on-demand versus prescheduled training have any effect on employee motivation to undertake the training? How do user experiences and gamification affect performance in Internet-based working environments (Thielsch & Niesenhaus, 2017)?

We know that poorly designed training will not stimulate and support learning, regardless of the extent to which appealing or expensive technology is used to deliver it (Brown & Ford, 2002; Kozlowski & Bell, 2003). Hence, if technology-based training is to be maximally effective, it must be designed to encourage active learning in participants. To do so, consider incorporating the following four principles into the instructional design (Brown & Ford, 2002):

1. Design the information structure and presentation to reflect both meaningful organization (or chunking) of material and ease of use.
2. Balance the need for learner control with guidance to help learners make better choices about content and process.
3. Provide opportunities for practice and constructive feedback.
4. Encourage learners to be mindful of their cognitive processing and in control of their learning processes.

Technique Selection

A training method can be effective only if it is used appropriately. Appropriate use, in this context, means rigid adherence to a two-step sequence: *first*, define what trainees are to learn, and *only then* choose a particular method that best fits these requirements. Far too often, unfortunately, trainers choose methods first and then force them to fit particular needs. This "retrofit" approach not only is wrong but also is often extremely wasteful of organizational resources—time, people, and money. It should be banished.

A technique is adequate to the extent that it provides the minimal conditions for effective learning to take place. To do this, a technique should do the following:

- Motivate the trainee to improve his or her performance
- Clearly illustrate desired skills
- Provide for the learner's active participation
- Provide an opportunity to practice
- Provide feedback on performance while the trainee learns
- Provide some means to reinforce the trainee while learning (e.g., using chatbots, automated yet personalized conversations between software and human users that may be used to provide reminders, track goals, assess transfer, and support continued performance; Han, 2017)
- Be structured from simple to complex tasks

- Be adaptable to specific problems
- Enable the trainee to transfer what is learned in training to other situations

Designers of training can apply this checklist to all proposed training techniques. If a particular technique appears to fit training requirements, yet is deficient in one or more areas, then either modify it to eliminate the deficiency or bolster it with another technique. The next step is to conduct the training. A checklist of the many logistical details involved is not appropriate here, but implementation should not be a major stumbling block if prior planning and design have been thorough. The final step, of course, is to measure the effects of training and their interaction with other organizational subsystems. To this topic, we now turn.

MEASURING TRAINING AND DEVELOPMENT OUTCOMES

"Evaluation" of a training program implies a dichotomous outcome (i.e., either a program has value or it does not). In practice, matters are rarely so simple, for outcomes are usually a matter of degree. To assess outcomes, we need to document systematically how trainees actually behave back on their jobs and the relevance of their behavior to the organization's objectives (Brown, 2017a; Machin, 2002; Snyder, Raben, & Farr, 1980). Beyond that, it is important to consider the intended purpose of the evaluation, as well as the needs and sophistication of the intended audience (Aguinis & Kraiger, 2009).

Why Measure Training Outcomes?

Evidence indicates that few companies assess the outcomes of training activities with any procedure more rigorous than participant reactions following the completion of training programs (Association for Talent Development, 2016; Brown, 2005; LinkedIn Learning, 2017; Sugrue & Rivera, 2005; Twitchell, Holton, & Trott, 2001). This is unfortunate because there are numerous reasons to evaluate training (Brown, 2017a; Noe, 2017; Sackett & Mullen, 1993):

- To make decisions about the future use of a training program or technique (e.g., continue, modify, eliminate)
- To compare the costs and benefits of training versus nontraining investments, such as work redesign or improved staffing
- To do a comparative analysis of the costs and benefits of alternative training programs
- To make decisions about individual trainees (e.g., certify as competent, provide additional training)
- To contribute to a scientific understanding of the training process
- To further political or public relations purposes (e.g., to increase the credibility and visibility of the training function by documenting success)

On a broader level, these reasons may be summarized as decision making, feedback, and marketing (Kraiger, 2002). Beyond these basic issues, we also would like to know whether the techniques used are more efficient or more cost effective than other available training methods.

Finally, we would like to be able to compare training with other approaches to developing workforce capability, such as improving staffing procedures and redesigning jobs. To do any of this, certain elements are essential.

Essential Elements of Measuring Training Outcomes

At the most basic level, the task of evaluation is counting—counting new customers, counting interactions, counting dollars, counting hours, and so forth. The most difficult tasks of evaluation are deciding *what* things to count and developing routine *methods* for counting them. As William Bryce Cameron (1963) famously said, "Not everything that counts can be counted, and not everything that can be counted counts" (p. 13). In the context of training, here is what counts (Campbell, Dunnette, Lawler, & Weick, 1970):

- Use of multiple criteria, not just for the sake of numbers, but also for the purpose of more adequately reflecting the multiple contributions of managers to the organization's goals.
- Some attempt to study the criteria themselves—that is, their relationships with each other and with other variables. The relationship between internal and external criteria is especially important.
- Enough experimental control to enable the causal arrow to be pointed at the training program. How much is enough will depend on the possibility of an interactive effect with the criterion measure and the susceptibility of the training program to the Hawthorne effect.
- Provision for saying something about the practical and theoretical significance of the results.
- A thorough, logical analysis of the process and content of the training.
- Some effort to deal with the "systems" aspects of training impact—that is, how training effects are altered by interaction with other organizational subsystems. For example, Kim and Ployhart (2014) used more than 12 years of longitudinal data to examine the effects of selective staffing and internal training on the financial performance of 359 firms during pre- and post-recessionary periods. They found a significant interaction between selective staffing and internal training, such that firms achieved consistent profit growth only when both were high.

Trainers must address these issues before they can conduct any truly meaningful evaluation of training impact. The remainder of this chapter treats each of these points more fully and provides practical illustrations of their use.

Criteria

As with any other HR program, the first step in judging the value of training is to specify multiple criteria. Although we covered the criterion problem already in Chapter 4, it is important to emphasize that the assessment of training outcomes requires multiple criteria because training is usually directed at specific components of performance. Organizations deal with multiple objectives, and training outcomes are multidimensional. Training may contribute to movement toward some objectives and away from others at the same time (Bass, 1983). Let's examine criteria according to time, type, and level.

Time

The important question here is "When, relative to the actual conduct of the training, should we obtain criterion data?" We could do so prior to, during, immediately after, or much later after the conclusion of training. To be sure, the timing of criterion measurement can make a great deal of difference in the interpretation of training effects (Sprangers & Hoogstraten, 1989). Thus, a study of 181 Korean workers (Lim & Morris, 2006) found that the relationship between perceived applicability (utility of training) and perceived application to the job (transfer) decreased as the time between training and measurement increased.

Conclusions drawn from an analysis of changes in trainees from before to immediately after training may differ drastically from conclusions based on the same criterion measures 6–12 months after training (Freeberg, 1976; Keil & Cortina, 2001; Steele-Johnson, Osburn, & Pieper, 2000). Yet both measurements are important. One review of 59 studies found, for example, that the time span of measurement (the time between the first and last observations) was one year or less for 26 studies, one to three years for 27 studies, and more than three years for only six studies (Nicholas & Katz, 1985). Comparisons of short- versus long-term training effects may yield valuable information concerning the interaction of training effects with other organizational processes (e.g., norms, values, leadership styles). Finally, it is not the absolute level of behavior (e.g., number of grievances per month, number of accidents) that is crucial, but rather the *change* in behavior from the beginning of training to some time after its conclusion.

Types of Criteria

It is important to distinguish internal from external criteria. Internal criteria are those that are linked directly to performance in the training situation. Examples of internal criteria are attitude scales and objective achievement examinations designed specifically to measure what the training program is designed to teach. External criteria, by contrast, are measures designed to assess actual changes in job behavior. For example, an organization may conduct a two-day training program in EEO law and its implications for talent management. A written exam at the conclusion of training (designed to assess mastery of the program's content) would be an internal criterion. Ratings by subordinates, peers, or supervisors and documented evidence regarding the trainees' on-the-job application of EEO principles constitute external criteria. Both internal and external criteria are necessary to evaluate the relative payoffs of training and development programs, and researchers need to understand the relationships among them in order to draw meaningful conclusions about training effects.

Criteria also may be qualitative or quantitative. Qualitative criteria are attitudinal and perceptual measures that usually are obtained by interviewing or observing employees or by administering written instruments. They are real-life examples of what quantitative results represent (Eden, 2017). Quantitative criteria also include measures of the outcomes of job behavior and system performance, which are often contained in employment, accounting, production, and sales records. These outcomes include turnover, absenteeism, dollar volume of sales, accident rates, and controllable rejects.

Both qualitative and quantitative criteria are important for a thorough understanding of training effects. Traditionally, researchers have preferred quantitative measures, except in organization development research (Austin & Bartunek, 2003; Nicholas, 1982; Nicholas & Katz, 1985). This may be a mistake, since there is much more to interpreting the outcomes of training than quantitative measures alone. By ignoring qualitative (process) measures, we may miss the richness of detail concerning *how* events occurred. Exclusive focus either on quantitative or qualitative measures, however, is short sighted and deficient. Thus, when learning and development (L&D) professionals were asked recently, "What are the top ways you measure the success of L&D at your company?" the five most common responses were qualitative and

the sixth had nothing to do with outcomes of a specific type of training per se. It was "length of time an employee stays at the company after completing a training" (LinkedIn Learning, 2017). At best, this offers an incomplete picture of the overall effects of training.

Finally, consider formative versus summative criteria. Formative criteria focus on evaluating training during program design and development, often through pilot testing. Based primarily on qualitative data such as opinions, beliefs, and feedback about a program from subject matter experts and sometimes customers, the purpose of formative evaluations is to make a program better. In contrast, the purpose of summative criteria is to determine if trainees have acquired the kinds of outcomes specified in training objectives. These may include knowledge, skills, attitudes, or new behaviors (Noe, 2017).

Levels of Criteria

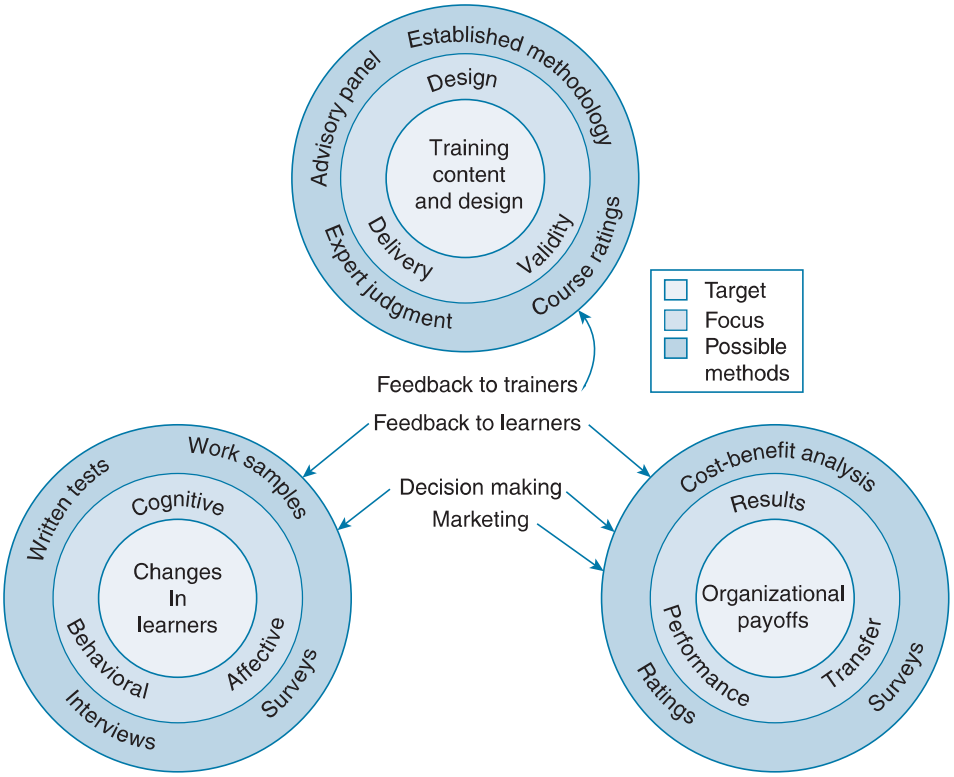
"Levels" of criteria may refer either to the organizational levels from which we collect criterion data or to the relative level of rigor we adopt in measuring training outcomes. With respect to organizational levels, information from trainers, trainees, subordinates, peers, supervisors, and the organization's policy makers (i.e., the training program's sponsors) can be extremely useful. In addition to individual sources, group sources (e.g., work units, teams, squads) can provide aggregate data regarding morale, turnover, grievances, and various cost, error, and/or profit measures that can be helpful in assessing training effects.

Kirkpatrick (1977, 1983, 1994) identified four levels of rigor in the evaluation of training and development programs: reaction, learning, behavior, and results. Note, however, that these levels provide only a vocabulary and a rough taxonomy for criteria. Higher levels do not necessarily provide more information than lower levels do, and the levels need not be causally linked or positively intercorrelated (Alliger & Janak, 1989). In general, there are four important concerns with Kirkpatrick's framework (Alliger, Tannenbaum, Bennett, Traver, & Shortland, 1997; Holton, 1996; Kraiger, 2002; Spitzer, 2005):

1. The framework is largely atheoretical; to the extent that it may be theory based, it is founded on an outdated behavioral perspective that ignores modern, cognitively based theories of learning.
2. It is overly simplistic in that it treats constructs such as trainee reactions and learning as unidimensional when, in fact, they are multidimensional (Alliger et al., 1997; Brown, 2005; Kraiger, Ford, & Salas, 1993; Morgan & Casper, 2001; Warr & Bunce, 1995). For example, reactions include affect toward the training as well as its perceived utility.
3. The framework makes assumptions about relationships between training outcomes that either are not supported by research (Bretz & Thompsett, 1992) or do not make sense intuitively. For example, Kirkpatrick argued that trainees cannot learn if they do not have positive reactions to the training. Yet a meta-analysis by Alliger et al. (1997) found an overall average correlation of only .07 between reactions of any type and immediate learning. In short, reactions to training should not be used blindly as a surrogate for the assessment of learning of training content.
4. Finally, the approach does not take into account the purposes for evaluation—decision making, feedback, and marketing (Kraiger, 2002).

Does Kirkpatrick's model suggest a causal chain across levels (positive reactions lead to learning, which leads to behavioral change, etc.), and do higher level evaluations provide the most informative data? Current thinking and evidence do not support these assumptions (Brown, 2017a). Rather, each level provides different, not necessarily better, information. Depending on the purpose of the evaluation, different outcomes will be more or less useful.

FIGURE 16.1 ■ An Integrative Model of Training Evaluation



Source: Republished with permission of John Wiley and Sons Inc., from Kraiger, K. (2002). Decision-based evaluation. In K. Kraiger (Ed.), *Creating, implementing, and managing effective training and development* (p. 343).

Figure 16.1 presents an alternative measurement model developed by Kraiger (2002), which attempts to overcome the deficiencies of Kirkpatrick's (1994) four-level model.

This approach clearly distinguishes evaluation targets (training content and design, changes in learners, and organizational payoffs) from data collection methods (e.g., with respect to organizational payoffs, cost-benefit analyses, ratings, and surveys). Targets and methods are linked through the options available for measurement—that is, its focus (e.g., with respect to changes in learners, the focus might be cognitive, affective, or behavioral changes). Finally, targets, focus, and methods are linked to evaluation purpose—feedback (to trainers or learners), decision making, and marketing. Kraiger (2002) also provided sample indicators for each of the three targets in Figure 16.1. For example, with respect to organizational payoffs, the focus might be on transfer of training (e.g., transfer climate, opportunity to perform, on-the-job behavior change), on results (performance effectiveness or tangible outcomes to a work group or organization), or on financial performance as a result of the training (e.g., through measures of return on investment or utility analysis) (Sung & Choi, 2014).

Additional Considerations in Measuring Training Outcomes

Regardless of the measures used, our goal is to be able to make meaningful inferences and to rule out alternative explanations for results. To do so, it is important to administer the measures according to some logical plan or procedure (experimental design) (e.g., before and

after training, as well as to a comparable control group). Numerous experimental designs are available for this purpose, and we consider them later in this chapter.

In assessing on-the-job behavioral changes, allow a reasonable period of time (e.g., at least three months) after the completion of training before taking measures. This is especially important for development programs that are designed to improve decision-making skills or to change attitudes or leadership styles. Such programs require *at least* three months before their effects manifest themselves in measurable behavioral changes. A large-scale meta-analysis reported an average interval of 133 days (almost 4.5 months) for the collection of outcome measures in behavioral terms (Arthur et al., 2003). To detect the changes, we need carefully developed techniques for systematic observation and measurement. Examples include scripted, job-related scenarios that use empirically derived scoring weights (Ostroff, 1991), behaviorally anchored rating scales, self-reports (supplemented by reports of subordinates, peers, and supervisors), critical incidents, or comparisons of trained behaviors with behaviors that were not trained (Frese, Beimel, & Schoenborn, 2003).

Strategies for Measuring Training Outcomes in Terms of Financial Impact

There continue to be calls for establishing the return on investment (ROI) for training, particularly as training activities continue to be outsourced and as new forms of technology-delivered instruction are marketed as cost effective (Association for Talent Development, 2016; LinkedIn Learning, 2017). Let's begin by examining what ROI is.

ROI relates program profits to invested capital. It does so in terms of a ratio in which the numerator expresses some measure of profit related to a project, and the denominator represents the initial investment in a program (Cascio, Boudreau, & Fink, in press). Suppose, for example, an organization invests $80,000 to design and deliver a wellness program. The program provides a total annual savings of $240,000 in terms of reduced sick days and improved health. The ROI is therefore [($240,000 – $80,000)/$80,000] × 100%, or 200%. Its net benefit per dollar spent is therefore 2:1. At a broader level, ROI has both advantages and disadvantages. Its major advantage is that it is simple and widely accepted. It blends in one number all the major ingredients of profitability, and it can be compared with other investment opportunities. On the other hand, it suffers from two major disadvantages. One, although the logic of ROI analysis appears straightforward, there is much subjectivity in determining the inflow of returns produced by an investment, how the inflows and outflows occur in each future time period, and how much what occurs in future time periods should be "discounted" to reflect greater risk and price inflation (Boudreau & Ramstad, 2006).

Two, typical ROI calculations focus on one HR investment at a time and fail to consider how those investments work together as a portfolio (Boudreau & Ramstad, 2007). Training may produce value beyond its cost, but would that value be even higher if it were combined with proper investments in individual incentives related to the training outcomes? As a general conclusion, ROI is best used when measurable outcomes are available (e.g., reductions in errors, sick days, or accidents), the training can be linked to an organizationwide strategy (e.g., cost reduction, improved customer service), it has management's interest, and it is attended by many employees (Noe, 2017).

Alternatively, financial outcomes may be assessed in terms of utility analysis (see Chapter 14). Such measurement is not easy, but the technology to do it is available and well developed. In fact, the basic formula for assessing the outcomes of training in dollar terms (Schmidt, Hunter, & Pearlman, 1982) builds directly on the general utility formula for assessing the payoff from selection programs (Equation 14.7):

$$\Delta U = (T)(N)(d_t)(SD_y) - (N)(C), \quad \textbf{(16.1)}$$

where ΔU is the dollar value of the training program, T is the number of years' duration of the training effect on performance, N is the number of persons trained, d_t is the true difference in job performance between the average trained worker and the average untrained worker in standard z-score units (see Equation 16.2), SD_y is the variability (standard deviation) of job performance in dollars of the untrained group, and C is the per-person cost of the training.

If the training is not held during working hours, then C should include only direct training costs. If training is held during working hours, then C should include, in addition to direct costs, all costs associated with having employees away from their jobs during the training. Employee time, for example, should include a full labor-cost multiplier (salary, benefits, and overhead). That value is a proxy for the opportunity costs of the lost value that employees or managers would be creating if they were not in training (Cascio et al., in press).

The term d_t is called the *effect size*. We begin with the assumption that there is no difference in job performance between trained workers (those in the experimental group) and untrained workers (those in the control group). The effect size tells us (*a*) if there is a difference between the two groups and (*b*) how large it is. The formula for effect size is

$$d_t = \frac{\bar{X}_e - \bar{X}_c}{SD\sqrt{r_{yy}}}, \quad \textbf{(16.2)}$$

where $\bar{X}_e$ is the average job performance of the trained workers (those in the experimental group), $\bar{X}_c$ is the average job performance of the untrained workers (those in the control group), SD is the standard deviation of the job performance measure in the untrained group, and r_{yy} is the reliability of the job performance measure (e.g., the degree of interrater agreement expressed as a correlation coefficient).

Equation 16.2 expresses effect size in standard-deviation units. To express it as a percentage, change in performance (X), the formula is

$$\%\text{ change in } X = d_t \times 100 \times SD_{pretest}/Mean_{pretest}, \quad \textbf{(16.3)}$$

where $100 \times SD_{pretest}/Mean_{pretest}$ (the coefficient of variation) is the ratio of the SD of pretest performance to its mean, multiplied by 100, where performance is measured on a ratio scale. Thus, to change d_t into a change-in-output measure, multiply d_t by the coefficient of variation for the job in question (Sackett, 1991).

When several studies are available, or when d_t must be estimated for a proposed human resource development (HRD) program, d_t is best estimated by the cumulated results of all available studies, using the methods of meta-analysis. Such studies are available in the literature (Arthur et al., 2003; Burke & Day, 1986; Guzzo, Jette, & Katzell, 1985; Morrow, Jarrett, & Rupinski, 1997). As they accumulate, managers will be able to rely on cumulative knowledge of the expected effect sizes associated with proposed HRD programs. Such a "menu" of effect sizes for HRD programs will allow HR professionals to compute the expected utilities of proposed HRD programs before the decision is made to allocate resources to such programs.

An Illustration of Utility Analysis

To illustrate the computation of the utility of training, suppose we wish to estimate the net payoff from a training program in supervisory skills. We develop the following information: $T = 2$ years; $N = 100$; $d_t = .31$ (Mathieu & Leonard, 1987); $SD_y = \$30{,}000$ (calculated by

any of the methods we discussed in Chapter 14); $C = \$4{,}000$ per person. According to Equation 16.1, the net payoff from the training program is

$$\Delta U = 2 \times 100 \times .31 \times \$30{,}000 - (100)\ (\$4{,}000)$$
$$\Delta U = \$1{,}460{,}000 \text{ over two years.}$$

Yet this figure is illusory because it fails to consider both economic and noneconomic factors that affect payoffs. For example, it fails to consider the fact that $1,460,000 received in two years is worth only $1,103,970 today (using the discount rate of 15% reported by Mathieu & Leonard, 1987). It also fails to consider the effects of variable costs and taxes (Boudreau, 1988). Finally, it looks only at a single cohort; but, if training is effective, managers want to apply it to multiple cohorts. Payoffs over subsequent time periods also must consider the effects of attrition of trained employees, as well as decay in the strength of the training effect over time (Cascio, 1989; Cascio et al., in press). Even after taking all of these considerations into account, the monetary payoff from training and development efforts still may be substantial and well worth demonstrating.

As an example, consider the results of a four-year investigation by a large, U.S.-based multinational firm of the effect and utility of 18 managerial and sales/technical training programs. The study is noteworthy, for it adopted a strategic focus by comparing the payoffs from different types of training in order to assist decision makers in allocating training budgets and specifying the types of employees to be trained (Morrow et al., 1997).

Over all 18 programs, the average improvement was about 17% (.54 SD). However, for technical/sales training, it was higher (.64 SD), and, for managerial training, it was lower (.31 SD). Thus, training in general was effective.

The mean ROI was 45% for the managerial training programs and 418% for the sales/technical training programs. However, one inexpensive time-management program developed in-house had an ROI of nearly 2,000%. When the economic utility of that program was removed, the overall average ROI of the remaining training programs was 84%, and the ROI of sales/technical training was 156%.

Why Not Hold All Training Programs Accountable Strictly in Economic Terms?

In practice, this is a rather narrow view of the problem, for economic indexes derived from the performance of operating units often are subject to bias (e.g., turnover, market fluctuations). Measures such as unit costs are not always under the exclusive control of the manager, and the biasing influences that are present are not always obvious enough to be compensated for.

This is not to imply that measures of results or financial impact should not be used to demonstrate a training program's worth; on the contrary, every effort should be made to do so. However, those responsible for assessing training outcomes should be well aware of the difficulties and limitations of measures of results or financial impact. They also must consider the utility of information-gathering efforts (i.e., if the costs of trying to decide whether the program was beneficial outweigh any possible benefits, then why make the effort?). At the same time, given the high payoff of effective management performance, the likelihood of such an occurrence is rather small. In short, don't ignore measures of results or financial impact. Thorough evaluation efforts consider measures of training content and design, measures of changes in learners, and organizational payoffs. Why? Because together they address each of the purposes of evaluation: to provide feedback to trainers and learners, to provide data on which to base decisions about programs, and to provide data to market them.

Influencing Managerial Decisions With Program-Evaluation Data

The real payoff from program-evaluation data is when the data lead to organizational decisions that are strategically important (Boudreau & Ramstad, 2007; Cascio et al., in press).

Mattson (2003) demonstrated convincingly that training-program evaluations that are expressed in terms of results do influence the decisions of operating managers to modify, eliminate, continue, or expand such programs. He showed that variables such as organizational cultural values (shared norms about important organizational values), the complexity of the information presented to decision makers, the credibility of that information, and the degree of its abstractness versus its concreteness affect managers' perceptions of the usefulness and ease of use of the evaluative information.

Other research has shed additional light on the best ways to present evaluation results to operating managers. To enhance managerial acceptance in the Morrow et al. (1997) study described earlier, the researchers presented the utility model and the procedures that they proposed to use to the CEO, as well as to senior strategic planning and HR managers, *before* conducting their research. They presented the model and procedures as fallible, but reasonable, estimates. As Morrow et al. (1997) noted, senior management's approval *prior* to actual application and consideration of utility results in a decision-making context is particularly important when one considers that nearly any field application of utility analysis will rely on an effect size calculated with an imperfect quasi-experimental design.

Mattson (2003) also recognized the importance of emphasizing the same things that managers of operating departments were paying attention to. Thus, in presenting results to managers of a business unit charged with sales and service, he emphasized outcomes attributed to the training program in terms that were important to those managers (volume of sales, employee-retention figures, and improvement in customer service levels). Clearly the "framing" of the message is critical and has a direct effect on its ultimate acceptability.

The sections that follow cover different types of experimental designs. This material is relevant and important for all readers, regardless of background. Even if you are not the person who conducts a study, but simply one who reads a report written by someone else, the discussion of experimental designs will help you to be a better informed, more critical, consumer of that information.

CLASSICAL EXPERIMENTAL DESIGNS

An experimental design is a plan, an outline for conceptualizing the relations among the variables of a research study. It also implies how to control the research situation and how to analyze the data (Kerlinger & Lee, 2000; Mitchell & Jolley, 2013). Experimental designs can be used with either internal or external criteria. For example, researchers can collect "before" measures on the job before training and collect "after" measures at the conclusion of training, as well as back on the job at some time after training. Researchers use experimental designs so that they can make causal inferences. That is, by ruling out alternative plausible explanations for observed changes in the outcome of interest, researchers want to be able to say that training caused the changes.

Unfortunately, most experimental designs and most training studies do not permit the causal arrow to point unequivocally toward training (x) as *the* explanation for observed results (y) (Eden, 2017). To do that, there are three necessary conditions (see Shadish, Cook, & Campbell, 2002, for more on this). The first requirement is that is that y did not occur until after x; the second is that x and y are actually shown to be related; and the third (and most difficult) is that other explanations of the relationship between x and y can be eliminated as plausible rival hypotheses.

To illustrate, consider a study by Batt (2002). The study examined the relationship among HR practices, employee quit rates, and organizational performance in the service sector. Quit rates were lower in establishments that emphasized high-involvement work systems. Batt (2002) showed that a range of HR practices was beneficial. Does that mean that the investments in training per se "caused" the changes in the quit rates and sales growth? No, but Batt (2002) did not claim that they did. Rather, she concluded that the entire set of HR practices *contributed* to the positive outcomes. It was impossible to identify the unique contribution of training alone. In fact, Shadish et al. (2002) suggest numerous potential contaminants or threats to valid interpretations of findings from field research. The threats may affect the following:

- *Statistical-conclusion validity*—the validity of inferences about the correlation (covariation) between treatment (e.g., training) and outcome
- *Internal validity*—the validity of inferences about whether changes in one variable caused changes in another
- *Construct validity*—the validity of inferences from the persons, settings, and cause-and-effect operations sampled within a study to the constructs these samples represent
- *External validity*—the validity of inferences about the extent to which results can be generalized across populations, settings, and times

In the context of training, let's consider 12 of these threats:

- *History*—specific events occurring between the "before" and "after" measurements in addition to training
- *Maturation*—ongoing processes within the individual, such as growing older or gaining job experience, which are a function of the passage of time
- *Testing*—the effect of a pretest on posttest performance
- *Instrumentation*—the degree to which an instrument may measure different attributes of an individual at two different points in time (e.g., parallel forms of an attitude questionnaire administered before and after training, or different raters rating behavior before and after training)
- *Statistical regression* (also known as regression to the mean)—changes in criterion scores resulting from selecting extreme groups on a pretest
- *Differential selection*—using different procedures to select individuals for experimental and control groups
- *Attrition*—differential loss of respondents from various groups
- *Interaction of differential selection and maturation*—that is, assuming experimental and control groups were different to begin with, the disparity between groups is compounded further by maturational changes occurring during the training period
- *Interaction of pretest with the experimental variable*—during the course of training, something reacts with the pretest in such a way that the pretest has a greater effect on the trained group than on the untrained group
- *Interaction of differential selection with training*—when more than one group is trained, differential selection implies that the groups are not equivalent on the criterion variable (e.g., skill in using a computer) to begin with; therefore, they may react differently to the training

- *Reactive effects of the research situation*—that is, the research design itself so changes the trainees' expectations and reactions that one cannot generalize results to future applications of the training
- *Multiple-treatment interference*—residual effects of previous training experiences affect trainees differently (e.g., finance managers and HR managers might not react comparably to a human relations training program because of differences in their previous training)

Table 16.1 presents examples of several experimental designs. These designs are by no means exhaustive; they merely illustrate the different kinds of inferences that researchers may draw and, therefore, underline the importance of considering experimental designs *before* training.

Design A

Design A, in which neither the experimental nor the control group receives a pretest, has not been used widely in training research. This is because the concept of the pretest is deeply ingrained in the thinking of researchers, although it is not essential to true experimental designs (Campbell & Stanley, 1963). We hesitate to give up "knowing for sure" that experimental and control groups were, in fact, "equal" before training, even though the most adequate all-purpose assurance of lack of initial biases between groups is randomization (Highhouse, 2009). Within the limits of confidence stated by tests of significance, randomization can suffice without the pretest (Campbell & Stanley, 1963, p. 25).

Design A controls for testing as main effect and interaction, but it does not measure them. Although such measurement is tangential to the real question of whether training did or did not produce an effect, the lack of pretest scores limits the ability to generalize, since it is impossible to examine the possible interaction of training with pretest ability level. In most organizational settings, however, variables such as job experience, age, or job performance are available either to use as covariates or to "block" subjects—that is, to group them in pairs matched on those variable(s) and then randomly to assign one member of each pair to the experimental group and the other to the control group. Both of these strategies increase statistical precision and make posttest differences more meaningful. In short, the main advantage of Design A is that it avoids pretest bias and the "give-away" repetition of identical or highly similar material (as in attitude-change studies), but this advantage is not without costs. For example, it does not prevent subjects from maturing or regressing; nor does it prevent events other than treatment (such as history) from occurring after the study begins (Shadish et al.,

TABLE 16.1 ■ Experimental Designs Assessing Training and Development Outcomes

	A		B	C		D			
	After-Only (One Control Group)		Before–After (No Control Group)	Before–After (One Control Group)		Solomon Four-Group Design Before–After (Three Control Groups)			
	E	C	E	E	C	E	C_1	C_2	C_3
Pretest	No	No	Yes	Yes	Yes	Yes	Yes	No	No
Training	Yes	No	Yes	Yes	No	Yes	No	Yes	No
Posttest	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: E refers to the experimental group. C refers to the control group.

2002). That said, when it is relatively costly to bring participants to an evaluation and administration costs are particularly high, after-only measurement of trained and untrained groups is best (Kraiger, McLinden, & Casper, 2004).

Design B

The defining characteristic of Design B is that it compares a group with itself. In theory, there is no better comparison, since all possible variables associated with characteristics of the subjects are controlled. In practice, however, when the objective is to measure change, Design B is fraught with difficulties, for numerous plausible rival hypotheses might explain changes in outcomes. History is one. If researchers administer pre- and posttests on different days, then events in between may have caused any difference in outcomes. Although the history effect is trivial if researchers administer pre- and posttests within a one- or two-hour period, it becomes more and more plausible as an alternative explanation for change as the time between pre- and posttests lengthens.

Aside from specific external events, various biological or psychological processes that vary systematically with time (i.e., maturation) also may account for observed differences. Hence, between pre- and posttests, trainees may have grown hungrier, more fatigued, or bored. "Changes" in outcomes simply may reflect these differences.

Moreover, the pretest itself may change that which is being measured. Hence, just the administration of an attitude questionnaire may change an individual's attitude; a manager who knows that his sales-meeting conduct is being observed and rated may change the way he behaves. In general, expect this reactive effect whenever the testing process is itself a stimulus to change rather than a passive record of behavior. The lesson is obvious: Use nonreactive measures whenever possible (cf. Rosnow & Rosenthal, 2008; Webb, Campbell, Schwartz, & Sechrest, 2000).

Instrumentation is yet a fourth uncontrolled rival hypothesis in Design B. If different raters do pre- and posttraining observation and rating, this could account for observed differences.

A fifth potential contaminant is statistical regression (i.e., less-than-perfect pretest–posttest correlations) (Furby, 1973; Kerlinger & Lee, 2000). This is a possibility whenever a researcher selects a group for training *because* of its extremity (e.g., all low scorers or all high scorers). Statistical regression has misled many a researcher time and again. The way it works is that lower scores on the pretest tend to be higher on the posttest and higher scores tend to be lower on the posttest when, in fact, no real change has taken place. This can deceive a researcher into concluding erroneously that a training program is effective (or ineffective). In fact, the higher and lower scores of the two groups may be due to the regression effect.

A control group allows one to "control" for the regression effect, since both the experimental and the control groups have pretest and posttest scores. If the training program has had a "real" effect, then it should be apparent over and above the regression effect. That is, both groups should be affected by the same regression and other influences, other things equal. So, if the groups differ in the posttest, it should be due to the training program (Kerlinger & Lee, 2000). The interaction effects (selection and maturation, testing and training, and selection and training) are likewise uncontrolled in Design B.

Despite all of the problems associated with Design B, it is still better to use it to assess change (together with a careful investigation into the plausibility of various threats), if that is the best one can do, than to do no evaluation. After all, organizations will make decisions about future training efforts with or without evaluation data (Kraiger et al., 2004; Sackett & Mullen, 1993). Moreover, if the objective is to measure individual achievement (a targeted level of performance), Design B can address that.

Design C

Design C (before–after measurement with a single control group) is adequate for most purposes, assuming that the experimental and control sessions are run simultaneously. Control is indispensable to the experimental method (Eden, 2017) and this design controls history,

maturation, and testing insofar as events that might produce a pretest–posttest difference for the experimental group should produce similar effects in the control group. We can control instrumentation either by assigning observers randomly to single sessions (when the number of observers is large) or by using each observer for both experimental and control sessions and ensuring that they do not know which subjects are receiving which treatments. Random assignment of individuals to treatments serves as an adequate control for regression or selection effects. Moreover, the data available for Design C enable a researcher to tell whether experimental mortality is a plausible explanation for pretest–posttest gain.

Information concerning interaction effects (involving training and some other variable) is important because, when present, interactions limit the ability to generalize results—for example, the effects of the training program may be specific only to those who have been "sensitized" by the pretest. In fact, when highly unusual test procedures (e.g., certain attitude questionnaires or personality measures) are used or when the testing procedure involves deception, surprise, stress, and the like, designs having groups that do not receive a pretest (e.g., Design A) are highly desirable, if not essential (Campbell & Stanley, 1963; Rosnow & Rosenthal, 2008). In general, however, *successful replication* of pretest–posttest changes at different times and in different settings increases our ability to generalize by making interactions of training with selection, maturation, instrumentation, history, and so forth less likely.

To compare experimental and control group results in Design C, either use analysis of covariance with pretest scores as the covariate, or analyze "change" scores for each group (Cascio & Kurtines, 1977; Cronbach & Furby, 1970; Edwards, 2002).

Design D

The most elegant of experimental designs, the Solomon (1949) four-group design (Design D), parallels Design C except that it includes two additional control groups (lacking the pretest). C_2 receives training plus a posttest; C_3 receives only a posttest. In this way, one can determine both the main effect of testing and the interaction of testing with training. The four-group design allows substantial increases in the ability to generalize, and, when training does produce changes in criterion performance, this effect is replicated in four different ways:

1. For the experimental group, posttest scores should be greater than pretest scores.
2. For the experimental group, posttest scores should be greater than C_1 posttest scores.
3. C_2 posttest scores should be greater than C_3 posttest scores.
4. C_2 posttest scores should be greater than C_1 pretest scores.

If data analysis confirms these directional hypotheses, this increases substantially the strength of inferences that can be drawn on the basis of this design. Moreover, by comparing C_3 posttest scores with experimental-group pretest scores and C_1 pretest scores, one can evaluate the combined effect of history and maturation.

Statistical analysis of the Solomon four-group design is not straightforward, since there is no one statistical procedure that makes use of all the data for all four groups simultaneously. Since all groups do not receive a pretest, the use of analysis of variance of gain scores (gain = posttest – pretest) is out of the question. Instead, consider a simple 2 × 2 analysis of variance of posttest scores (Solomon, 1949):

	No Training	Training
Pretested	C_1	E
Not Pretested	C_3	C_2

Estimate training main effects from column means, estimate pretesting main effects from row means, and estimate interactions of testing with training from cell means.

Despite its apparent advantages, the Solomon four-group design is not without theoretical and practical problems (Bond, 1973; Kerlinger & Lee, 2000). For example, it assumes that the simple passage of time and training experiences affect all posttest scores independently. However, some interaction between these two factors is inevitable, thus jeopardizing the significance of comparisons between posttest scores for C_3 and pretest scores for E and C_1.

Serious practical problems also may emerge. The design requires large numbers of persons in order to represent each group adequately and to generate adequate statistical power. For example, in order to have 30 individuals in each group, the design requires 120 participants. This may be impractical or unrealistic in many settings.

Here is a practical example of these constraints (Sprangers & Hoogstraten, 1989). In two field studies of the impact of pretesting on posttest responses, the researchers used nonrandom assignment of 37 and 58 subjects in a Solomon four-group design. Their trade-off of low statistical power for greater experimental rigor illustrates the extreme difficulty of applying this design in field settings.

A final difficulty lies in the application of the four-group design. Solomon (1949) has suggested that, after the value of the training is established using the four groups, the two control groups that did not receive training then could be trained, and two new groups could be selected to act as controls. In effect, this would replicate the entire study—but would it? Sound experimentation requires that conditions remain constant, but it is quite possible that the first training program may have changed the organization in some way, so that those who enter the second training session already have been influenced.

Cascio (1976a) showed this empirically in an investigation of the stability of factor structures in the measurement of attitudes. The factor structure of a survey instrument designed to provide a baseline measure of managerial attitudes toward African Americans in the working environment did not remain constant when compared across three different samples of managers from the same company at three different time periods. During the two-year period that the training program ran, increased societal awareness of EEO, top management emphasis of it, and the fact that over 2,200 managers completed the training program probably altered participants' attitudes and expectations even before the training began.

Despite its limitations, when it is possible to apply the Solomon four-group design realistically, to assign subjects randomly to the four groups, and to maintain proper controls, this design controls most of the sources of invalidity that it is possible to control in one experimental design. Table 16.2 presents a summary of the sources of invalidity for Designs A through D.

Limitations of Experimental Designs

Having illustrated some of the nuances of experimental design, let's pause for a moment to place design in its proper perspective. First, exclusive emphasis on the design aspects of measuring training outcomes is rather narrow in scope. An experiment usually settles on a single criterion dimension, and the whole effort depends on observations of that dimension (Newstrom, 1978; Weiss & Rein, 1970). Hence, experimental designs are quite limited in the amount of information they can provide. There is no logical reason for investigators to consider just a single criterion dimension, but this is usually what happens. Ideally, an experiment should be part of a continuous feedback process rather than just an isolated event or demonstration (Shadish et al., 2002; Snyder et al., 1980).

TABLE 16.2 ■ Source of Invalidity for Experimental Designs A Through D

Design \ Source	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction of Selection and Maturation	Interaction of Testing and Training	Interaction of Selection and Training	Reactive Arrangements	Multiple-Treatment Interference
A. After–only (one control)	+	+	+	+	+	+	+	+	+	?	?	
B. Before–after (no control)	–	–	–	–	?	+	+	–	–	–	?	
C. Before–after (one control)	+	+	+	+	+	+	+	+	+	?	?	
D. Before–after (three controls) Solomon four–group design		+	+	+	+	+	+	+	+	+	?	?

Note: A "+" indicates that the factor is controlled, a "–" indicates that the factor is not controlled, a "?" indicates possible source of concern, and a blank indicates that the factor is not relevant. See text for appropriate qualifications regarding each design.

Second, meta-analytic reviews have demonstrated that effect sizes obtained from single-group pretest–posttest designs (Design B) are systematically higher than those obtained from control or comparison-group designs (Carlson & Schmidt, 1999; Lipsey & Wilson, 1993). Type of experimental design therefore moderates conclusions about the effectiveness of training programs. Fortunately, corrections to mean effect sizes for data subgrouped by type of dependent variable (differences are most pronounced when the dependent variable is knowledge assessment) and type of experimental design can account for most such biasing effects (Carlson & Schmidt, 1999).

Third, it is important to ensure that any attempt to measure training outcomes through the use of an experimental design has adequate statistical power. Power is the probability of correctly rejecting a null hypothesis when it is false (Murphy & Myors, 2003). Research indicates that the power of training-evaluation designs is a complex issue, for it depends on the effect size obtained, the reliability of the dependent measure, the correlation between pre- and posttest scores, the sample size, and the type of design used (Arvey, Cole, Hazucha, & Hartanto, 1985). Software that enables straightforward computation of statistical power and confidence intervals (*Power & Precision*, 2000) should make power analysis a routine component of training-evaluation efforts.

Finally, experiments often fail to focus on the real goals of an organization. For example, experimental results may indicate that job performance after treatment A is superior to performance after treatment B or C. The really important question, however, may not be whether treatment A is more effective, but rather what levels of performance we can expect from almost all trainees at an acceptable cost and the extent to which improved performance through training "fits" the broader strategic thrust of an organization. Box 16.1 is a practical illustration of a true field experiment.

BOX 16.1

PRACTICAL ILLUSTRATION: A TRUE FIELD EXPERIMENT WITH A SURPRISE ENDING

The command teams of 18 logistics units in the Israel Defense Forces were assigned randomly to experimental and control conditions. Each command team included the commanding officer of the unit plus subordinate officers, both commissioned and noncommissioned. The command teams of the nine experimental units underwent an intensive three-day team-development workshop. The null hypothesis was that the workshops had no effect on team or organizational functioning (Eden, 1985).

The experimental design provided for three different tests of the hypothesis, in ascending order of rigor. First, a Workshop Evaluation Questionnaire was administered to team members after the workshop to evaluate their subjective *reactions* to its effectiveness.

Second, Eden (1985) assessed the before-and-after perceptions of command team members in both the experimental and the control groups by means of a Team Development Questionnaire, which included ratings of the team leader, subordinates, team functioning, and team efficiency. This is a true experimental design (Design C), but its major weakness is that the outcomes of interest were assessed in terms of responses from team members who personally had participated in the workshops. This might well lead to positive biases in the responses.

To overcome this problem, Eden used a third design. He selected at random about 50 subordinates representing each experimental and control unit to complete the Survey of Organizations both before and after the team-development workshops. This instrument measures organizational functioning in terms of general management, leadership, coordination, three-way communications, peer relations, and satisfaction. Since subordinates had no knowledge of the team-development workshops and therefore no ego involvement in them, this design represents the most internally valid test of the hypothesis. Moreover, since an average of 86% of the subordinates drawn from the experimental-group units completed the posttraining questionnaires, as did an average of 81% of those representing control groups, Eden could rule out the effect of attrition as a threat to the internal validity of the experiment. Rejection of the null hypothesis would imply that the effects of the team-development effort really did affect the rest of the organization.

To summarize: Comparison of the command team's before-and-after perceptions tests whether the workshop influenced the team; comparison of the subordinates' before-and-after perceptions tests whether team development affected the organization. In all, 147 command-team members and 600 subordinates completed usable questionnaires.

Results

Here's the surprise: Only the weakest test of the hypothesis, the postworkshop reactions of participants, indicated that the training was effective. Neither of the two before-and-after comparisons detected any effects, either on the team or on the organization. Eden (1985) concluded:

> The safest conclusion is that the intervention had no impact. This disconfirmation by the true experimental designs bares the frivolity of self-reported after-only perceptions of change. Rosy testimonials by [trainees] may be self-serving, and their validity is therefore suspect. (p. 98)

QUASI-EXPERIMENTAL DESIGNS

In field settings, there often are major obstacles to conducting true experiments. True experiments require the manipulation of at least one independent variable, the random assignment of participants to groups, and the random assignment of treatments to groups (Kerlinger & Lee, 2000). Managers may disapprove of the random assignment of people to conditions. Line managers do not see their subordinates as interchangeable, like pawns on a chessboard, and they often distrust randomness in experimental design. Beyond that, some managers see training evaluation as disruptive and expensive. Eden (2017) offered eight strategies for overcoming deterrents to field experimentation, including the avoidance of jargon, explaining

randomization to lay managers, transforming proprietary data, and using emerging technologies, such as experience sampling (Beal, 2015).

Despite calls for more rigor in training-evaluation designs (Littrell, Salas, Hess, Paley, & Riedel, 2006; Shadish & Cook, 2009; Wang, 2002), some less-complete (i.e., quasi-experimental) designs can provide useful data even though a true experiment is not possible (Grant & Wall, 2009). What makes them "quasi" is their lack of randomly created, preexperimental equivalence, which degrades internal validity (Eden, 2017). Shadish et al. (2002) offered a number of quasi-experimental designs with the following rationale: The central purpose of an experiment is to eliminate alternative hypotheses that also might explain results. If a quasi-experimental design can help eliminate some of these rival hypotheses, then it may be worth the effort.

Because full experimental control is lacking in quasi-experiments, it is important to know which specific variables are uncontrolled in a particular design (cf. Tables 16.2 and 16.3). Investigators should, of course, design the very best experiment possible, given their circumstances, but where full control is not possible, they should use the most rigorous design that *is* possible. For these reasons, we present four quasi-experimental designs, together with their respective sources of invalidity, in Table 16.3.

TABLE 16.3 ■ Source of Invalidity for Four Quasi–Experimental Designs

Design \ Source	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction of Selection and Maturation	Interaction of Testing and Training	Interaction of Selection and Training	Reactive Arrangements	Multiple-Treatment Interference
E. Time-series design Measure (M) M (Train) MMM	–	+	+	?	+	+	+	+	–	?	?	
F. Nonequivalent control-group design I. M train M II. M no train M	+	+	+	+	?	+	+	–	–	?	?	
G. Nonequivalent dependent variable design I. M (experimental and control variables) train II. M (experimental and control variables)	–	–	+	?	–	+	+	+	+	+	+	
H. Institutional cycle design Time 1 2 3 I. M (train) M (no train) M II. M (no train) M (train) M	+	–	+	+	?	–	?		+	?	+	

Note: A "+" indicates that the factor is controlled, a "–" indicates that the factor is not controlled, a "?" indicates a possible source of concern, and a blank indicates that the factor is not relevant.

Design E

The *time-series design* is especially relevant for assessing the outcomes of training and development programs. It uses a single group of individuals and requires that criterion data be collected at several points in time, both before and after training. Criterion measures obtained before the introduction of the training experience then are compared to those obtained after training. A curve relating criterion scores to time periods may be plotted, and, in order for an effect to be demonstrated, there should be a discontinuity or change in the series of measures, corresponding to the training program, that does not occur at any other point. This discontinuity may represent an abrupt change either in the slope or in the intercept of the curve. The time-series design is frequently used to evaluate training programs that focus on improving readily observable outcomes, such as accident rates, productivity, and absenteeism. By incorporating a large number of observations pre- and posttraining, it allows researchers to analyze the stability of training outcomes over time. To rule out alternative explanations for evaluation results, consider using comparison groups or reversal (a time period where participants no longer receive the intervention) (Noe, 2017).

Design F

Another makeshift experimental design, Design F, is the *nonequivalent control-group design*. Although Design F appears identical to Design C (before–after measurement with one control group), there is a critical difference: In Design F, individuals from a common population are not assigned randomly to the experimental and control groups. This design is common in applied settings where naturally occurring groups must be used (e.g., work group A and work group B). Design F is especially appropriate when Designs A and C are impossible because even the addition of a nonequivalent control group makes interpretation of the results much less ambiguous than in Design B, the single-group pretest–posttest design. Needless to say, the nonequivalent control group becomes much more effective as an experimental control as the similarity between experimental and control-group pretest scores increases. Box 16.2 illustrates the hazards of nonequivalent designs.

The major sources of invalidity in this design are the selection-maturation interaction and the testing-training interaction. For example, if the experimental group happens to consist of young, inexperienced workers and the control group consists of older, highly experienced workers who are tested and retested, a gain in criterion scores that appears specific to the experimental group might well be attributed to the effects of training when, in fact, the gain would have occurred even without training.

Regression effects pose a further threat to unambiguous inferences in Design F. This is certainly the case when experimental and control groups are "matched" (which is no substitute

BOX 16.2

PRACTICAL ILLUSTRATION: THE HAZARDS OF NONEQUIVALENT DESIGNS

The hazards of nonequivalent designs are illustrated neatly in the evaluations of a training program designed to improve the quality of group decisions by increasing the decision-making capabilities of its members. A study by Bottger and Yetton (1987) that demonstrated the effectiveness of this approach used experimental and control groups whose pretest scores differed significantly. When Ganster, Williams, and Poppler (1991) replicated the study using a true experimental design (Design C) with random assignment of subjects to groups, the effect disappeared.

for randomization), yet the pretest means of the two groups differ substantially. When this happens, changes in criterion scores from pretest to posttest may well be due to regression effects, not training. Despite these potential contaminants, we encourage increased use of Design F, especially in applied settings. Be aware of potential contaminants that might make results equivocal, and attempt to control them as much as possible. That said, do not assume that statistical control after the experiment has been conducted can substitute for random assignment to treatments (Carlson & Wu, 2012).

Design G

We noted earlier that many managers reject the notion of random assignment of participants to training and no-training (control) groups. A type of design that those same managers may find useful is the nonequivalent dependent variable design (Shadish et al., 2002) or "internal referencing" strategy (Haccoun & Hamtieux, 1994). The design is based on a single treatment group and compares two sets of dependent variables—one that training should affect (experimental variables), and the other that training should not affect (control variables). Design G can be used whenever the evaluation is based on some kind of performance test.

Perhaps the major advantage of this design is that it effectively controls two important threats to internal validity: testing and the Hawthorne effect (i.e., simply reflecting on one's behavior as a result of participating in training could produce changes in behavior). Another advantage, especially over a nonequivalent control-group design (Design F), is that there is no danger that an unmeasured variable that differentiates the nonequivalent control group from the trained group might interact with the training. For example, it is possible that self-efficacy might be higher in the nonequivalent control group because volunteers for such a control group may perceive that they do not need the training in question (Frese et al., 2003).

Design G does not control for history, maturation, and regression effects, but its most serious potential disadvantage is that the researcher is able to control how difficult or easy it is to generate significant differences between the experimental and control variables. The researcher can do this by choosing variables that are very different from or similar to those that are trained.

To avoid this problem, choose control variables that are conceptually similar to, but distinct from, those that are trained. For example, in a program designed to teach inspirational communication of a vision as part of training in charismatic leadership, Frese et al. (2003) included the following as part of set of experimental (trained) items: variation of speed, variation of loudness, and use of "we." Control (untrained) items included, among others, the following: combines serious/factual information with witty and comical, examples from practice, and good organization, such as a, b, and c. The control items were taken from descriptions of two training seminars on presentation techniques. A different group of researchers independently coded them for similarity to inspirational speech, and the researchers chose items coded to be least similar.

Before–after coding of behavioral data indicated that participants improved much more on the trained variables than on the untrained variables (effect sizes of about 1.0 versus .3). This suggests that training worked to improve the targeted behaviors but did not systematically influence the untargeted behaviors. At the same time, we do not know if there were long-term, objective effects of the training on organizational performance or on the commitment of subordinates.

Design H

A final quasi-experimental design, appropriate for cyclical training programs, is known as the *recurrent institutional cycle design*. It is Design H in Table 16.3. For example, a large sales

organization presented a management development program, known as the State Manager Program, every two months to small groups (12–15) of middle managers (state managers). The one-week program focused on all aspects of retail sales (e.g., new product development, production, distribution, marketing, merchandising). The program was scheduled so that all state managers (approximately 110) could be trained over an 18-month period. This is precisely the type of situation for which Design H is appropriate—that is, a large number of persons will be trained, but not all at the same time. Different *cohorts* are involved. Design H is actually a combination of two (or more) before–after studies that occur at different points in time. Group I receives a pretest at time 1, then training, and then a posttest at time 2. At the same chronological time (time 2), Group II receives a pretest, training, and then a posttest at time 3. At time 2, therefore, an experimental and a control group have, in effect, been created. One can obtain even more information (and with quasi-experimental designs, it is *always* wise to collect as much data as possible or to demonstrate the effect of training in several different ways) if it is possible to measure Group I again at time 3 and to give Group II a pretest at time 1. This controls the effects of history. Moreover, the time 3 data for Groups I and II and the posttests for all groups trained subsequently provide information as to how the training program is interacting with other organizational events to produce changes in the criterion measure.

Several cross-sectional comparisons are possible with the cycle design:

- Group I posttest scores at time 2 can be compared with Group II pretest scores at time 2.
- Gains made in training for Group I (time 2 posttest scores) can be compared with gains in training for Group II (time 3 posttest scores).
- Group II posttest scores at time 3 can be compared with Group I posttest scores at time 3 (i.e., gains in training versus gains [or no gains] during the no-training period).

This design controls history and test–retest effects but not differences in selection. One way to control for possible differences in selection, however, is to split one of the groups (assuming it is large enough) into two equated samples, one measured both before and after training and the other measured only after training:

	Time 2	Time 3	Time 4
Group II_a	Measure	Train	Measure
Group II_b	—	Train	Measure

Comparison of the posttest scores of two carefully equated groups (Groups II_a and II_b) is more precise than a similar comparison of posttest scores of two unequated groups (Groups I and II).

A final deficiency in the cycle design is the lack of adequate control for the effects of maturation. This is not a serious limitation if the training program is teaching specialized skills or competencies, but it is a plausible rival hypothesis when the objective of the training program is to change attitudes.

Campbell and Stanley (1963) expressed aptly the logic of these makeshift designs:

> [O]ne starts out with an inadequate design and then adds specific features to control for one or another of the recurrent sources of invalidity. The result is often an inelegant

> accumulation of precautionary checks, which lacks the intrinsic symmetry of the "true" experimental designs, but nonetheless approaches experimentation. (p. 57)

Other quasi-experimental designs (cf. Grant & Wall, 2009; Kerlinger & Lee, 2000; Shadish et al., 2002) are appropriate in specialized situations, but the ones we have discussed seem well suited to the types of problems that applied researchers are likely to encounter.

STATISTICAL, PRACTICAL, AND THEORETICAL SIGNIFICANCE

As in selection, the problem of *statistical* versus *practical* significance is relevant for the assessment of training outcomes. Demonstrations of statistically significant change scores may mean little in a practical sense. From a practical perspective, researchers must show that the effects of training *do* make a difference to organizational goals—in terms of lowered production costs, increased sales, fewer grievances, and so on. Practical significance typically is reflected in terms of effect sizes or measures of variance accounted for (Grissom & Kim, 2014; Schmidt & Hunter, 2014).

A related issue concerns the relationship between practical and theoretical significance. Training researchers frequently are content to demonstrate only that a particular program "works"—the prime concern being to sell the idea to top management or to legitimize an existing (perhaps substantial) investment in a particular development program. This is only half the story. The real test is whether the new training program is superior to previous or existing methods for accomplishing the same objectives. To show this, firms need systematic research to evaluate the effects of independent variables that are likely to affect training outcomes—for example, different training methods, different depths of training, or different types of media for presenting training.

If researchers adopt this two-pronged approach to measuring training outcomes and if they can map the effects of relevant independent variables across different populations of trainees and across different criteria, then the assessment takes on theoretical significance. For example, using meta-analysis, Arthur et al. (2003) found medium-to-large effect sizes for organizational training (sample-weighted average effect sizes of .60 for reaction criteria, .63 for measures of learning, and .62 for measures of behavior or results). Other organizations and other investigators may use this knowledge to advantage in planning their own programs. The concept of statistical significance, while not trivial, in no sense guarantees practical or theoretical significance—the major issues in outcome measurement.

LOGICAL ANALYSIS

Experimental control is but one strategy for responding to criticisms of the internal or statistical conclusion validity of a research design (Eden, 2017; McLinden, 1995; Sackett & Mullen, 1993). A logical analysis of the process and content of training programs can further enhance our understanding of *why* we obtained the results we did. As we noted earlier, both qualitative and quantitative criteria are important for a thorough understanding of training effects. Here are some qualitative issues to consider:

- Were the goals of the training clear both to the organization and to the trainees?
- Were the methods and content of the training relevant to the goals?

- Were the proposed methods used and the proposed content taught?
- Did it appear that learning was taking place?
- Does the training program conflict with any other program in the organization?
- What kinds of criteria should be expected to show change as a result of the training?

For every one of these questions, supplement the subjective opinions of experts with objective data. For example, to provide broader information regarding the second question, document the linkage between training content and job content. A quantitative method is available for doing this (Bownas, Bosshardt, & Donnelly, 1985). It generates a list of tasks that receive undue emphasis in training, those that are not being trained, and those that instructors intend to train but that graduates report being unable to perform. It proceeds as follows:

1. Identify curriculum elements in the training program.
2. Identify tasks performed on the job.
3. Obtain ratings of the emphasis given to each task in training, of how well it was learned, and of its corresponding importance on the job.
4. Correlate the two sets of ratings—training emphasis and job requirements—to arrive at an overall index of fit between training and job content.
5. Use the ratings of training effectiveness to identify tasks that appear to be over- or underemphasized in training.

Confront these kinds of questions during program planning *and* evaluation. When integrated with responses to the other issues presented earlier in this chapter, especially the "systems" aspects of training impact, training outcomes become much more meaningful. This is the ultimate payoff of the measurement effort.

In Chapter 17, we continue our presentation by examining emerging international issues in applied psychology and talent management. We begin by considering the growth of HR management issues across borders.

EVIDENCE-BASED IMPLICATIONS FOR PRACTICE

- Numerous training methods and techniques are available, but each one can be effective only if it is used appropriately. To do that, first define what trainees are to learn, and only then choose a particular method that best fits these requirements.
- In evaluating training outcomes, be clear about your purpose. Three general purposes are to provide feedback to trainers and learners, to provide data on which to base decisions about programs, and to provide data to market them.
- Use quantitative as well as qualitative measures of training outcomes. Each provides useful information.
- Regardless of the measures used, the overall goal is to be able to make meaningful inferences and to rule out alternative explanations for results. To do that, it is important to administer the measures according to some logical plan or procedure (experimental or quasi-experimental design). Be clear about what threats to valid inference your design controls for and fails to control for.
- No less important is a logical analysis of the process and content of training programs, for it can enhance understanding of *why* we obtained the results we did.

Discussion Questions

1. Discuss two methods of training that illustrate each of the following categories: presentation, hands-on, group building, and technology based.
2. Discuss the advantages and disadvantages of interactive multimedia training.
3. Why are both quantitative and qualitative criteria important to consider in the assessment of training outcomes?
4. What criteria would you use to select one training method over another?
5. When does it make the most sense to use ROI to assess training outcomes?
6. Describe some of the key differences between experimental and quasi-experimental designs.
7. Your boss asks you to design a study to evaluate the effects of a training class in stress reduction. How will you proceed?
8. Your firm decides to train its entire population of employees and managers (500) to provide "legendary customer service." Suggest a design for evaluating the impact of such a massive training effort.
9. Identify two examples that illustrate the difference between statistical and practical significance.
10. What additional information might a logical analysis of training outcomes provide that an experimental or quasi-experimental design does not?